UNIVERSIDAD INTERNACIONAL DE LA RIOJA

Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos

Detección de Didymosphenia Geminata en cuerpos de agua dulce utilizando técnicas de Inteligencia Artificial

Trabajo Fin de Máster

Tipo de trabajo: Comparativa de soluciones

Presentado por: Aracena Vallejos, Estefania – Medina Jiménez, Lenin

Directora: Prados Privado, María

Resumen

El presente trabajo fin de máster aborda el problema de la detección y el control de *Didymosphenia geminata* (Didymo), una especie de alga invasora que se ha extendido por varios países, incluyendo Chile, donde ha causado graves daños a los ecosistemas acuáticos y a las actividades humanas relacionadas con el agua. El objetivo principal es desarrollar un sistema de alerta basado en técnicas de inteligencia artificial que permita identificar la presencia de Didymo en distintas subcuencas hidrográficas y estimar el grado de cobertura algal. Para ello, se utiliza un conjunto de datos proporcionado por el Instituto de Fomento Pesquero de Puerto Montt, que contiene información sobre parámetros físicos y químicos del agua, así como sobre la presencia o ausencia de Didymo y el porcentaje de cobertura algal en cada muestra. Se aplican diferentes métodos de análisis de datos, tales como *clustering*, *ensembled learning* (*random forest*, redes neuronales) y redes convolucionales para grafos, para crear modelos de clasificación y predicción que se comparan entre sí en términos de precisión y robustez. Los resultados obtenidos muestran que las redes convolucionales para grafos son las más adecuadas para este problema, ya que son capaces de capturar mejor las relaciones entre los parámetros y las características de cada sitio de muestreo con el porcentaje de cobertura de Didymo en las subcuencas de los ríos del sur de Chile.

Palabras Clave: *Didymosphenia geminata*, Didymo, inteligencia artificial, *clustering*, *ensembled learning*, grafos, redes convolucionales

Abstract

This master's thesis follows the detection and control issue of Didymosphenia geminata (Didymo), an invasive algae species that has been spread in many countries, including Chile, and it has caused serious damage to aquatic ecosystems and a negative impact in activities related to water. The aim is to develop an alert system based on artificial intelligence techniques that allows to identify the presence of Didymo within different hydrographic subbasins and rate the algae coverage. To complete the task, a data set provided by the Instituto Fomento Pesquero from Puerto Montt has been handled, it includes information about physical and chemical parameters of the water as well as the presence or absence of Didymo and the algae coverage percentage in each sample. Different methods of data analysis have been applied, such as clustering, ensemble learning (random forest, neural networks) and graph convolutional networks, to develop classification and prediction models that are compare with each other in terms of accuracy and robustness. The results showed that graph convolutional networks are the most suitable method for this problem, they are capable to capture the relationships between the parameters and the characteristics of each sampling site with the coverage percentage of Didymo at the subbasins of the chilean southern rivers.

Keywords: Didymosphenia geminata, Didymo, artificial intelligence, clustering, ensembled learning, graph, convolutional networks

Agradecimientos

Agradecimientos al equipo de trabajo del "Monitoreo, prospección e investigación de la especie plaga *Didymosphenia geminata* en ecosistemas fluviales y lacustres de la zona centro, sur y austral de Chile" que es ejecutado por el Departamento de Medio Ambiente de la División de Investigación en Acuicultura del Instituto de Fomento Pesquero (sede Puerto Montt), que es parte del programa de investigación permanente de la Subsecretaría de Pesca y Acuicultura, financiado por el Ministerio de Economía.

También, agradecimientos a Renzo Valencia, quién nos quién nos guio a los grafos y logramos obtener resultados exitosos.

Agradecemos a la Universidad Internacional de La Rioja (UNIR) y a sus profesores por ofrecernos una formación de calidad y adaptada a nuestras necesidades. Su metodología online nos permitió compatibilizar nuestros estudios con nuestra actividad profesional y personal. Su rigor académico y su compromiso con la excelencia nos han enriquecido tanto a nivel intelectual como humano. De igual manera, agradecemos a nuestra tutora del TFM, la Doctora María Pérez, por su dedicación, paciencia y sabiduría. Su asesoramiento, feedback y sugerencias fueron clave para mejorar la calidad de mi trabajo y superar las dificultades que se presentaron. Su confianza nos motivó a esforzarnos al máximo y a aprender de cada etapa del proceso.

Estefania Aracena & Lenin Medina

Agradezco a mi padre Manuel Ignacio Aracena Henríquez y mi madre Orietta Bertides Vallejos Garay por acompañarme en cada aventura y darme un empujón ante las dudas, y creer y confiar en que cada día puedo ser mejor que el día anterior. A mi hermano Marcelo y Kuky, por aguantarme al invadir sus espacios. A Alejandra Oyanedel, quién aceptó esta loca idea de trabajar con los datos que recopiló con su equipo y a mi compañero Lenin Medina, con quien formamos un buen equipo y logramos culminar esta etapa estudiantil.

Estefania Aracena

Quiero expresar mi profundo agradecimiento a las personas que más quiero y que me han acompañado en este camino. A mi esposa Carina Baquero, que me alentó y apoyó a obtener la maestría. Su amor, comprensión y paciencia fueron esenciales para que pudiera cumplir este sueño. Gracias por estar siempre a mi lado y por creer en mí.

A Mateo y Felipe mis hijos, quienes son mi mayor inspiración. Ellos son mi orgullo y mi alegría. Gracias por su cariño, respeto y admiración. Espero que este logro les inspire a seguir sus propias metas y a nunca rendirse.

A mi madre Elsa y mi hermano Raúl, su confianza y apoyo incondicional fueron muy valiosos para mí. Su generosidad, solidaridad y lealtad han sido invaluables para mí. Los quiero mucho y les estoy muy agradecido.

A mi compañera Estefanía Aracena, quien contagió su disciplina y búsqueda de la excelencia para alcanzar este reto.

Lenin Medina

Tabla de participación

Apartado de la memoria	Responsables
Resumen	Lenin Medina
Introducción	Estefania Aracena y Lenin Medina
Contexto y estado del arte	Estefania Aracena y Lenin Medina
Objetivos concretos y metodología de trabajo	Estefania Aracena y Lenin Medina
Desarrollo específico de la contribución	Estefania Aracena y Lenin Medina
Conclusiones y trabajo futuro	Estefania Aracena
Código implementado	Estefania Aracena y Lenin Medina
Formato de memoria Trabajo Fin de Máster	Estefania Aracena
Formato códigos implementados	Lenin Medina

Índice de contenidos

Índice de tablas

Índice de figuras

1. Introducción

Desde el año 2010, Chile ha combatido el alga *Didymosphenia geminata*, conocida como Didymo, una plaga que acecha sistemas hidrográficos como ríos y lagos en todo el mundo. Originaría del hemisferio norte, esta se ha distribuido a lo largo de la zona central y sur de Chile, generando efectos negativos, tanto económicos como ecológicos. Esta alga tiene la particularidad de adherirse a superficies, como rocas, y al reproducirse, forma mantos compactos cubriendo grandes superficies acuáticas, evitando que la flora y fauna del lugar invadido se desarrolle normalmente, esto debido baja disponibilidad de nutrientes y a la nula disposición de luz natural causadas por el alga (Hix & Murdock, 2019; Salvo & Oyanedel, 2019). Si bien no afecta a la salud humana, zonas de interés turístico se han visto afectadas ante la presencia de densas colonias de esta microalga, reduciendo el flujo de turistas que realizan pesca deportiva e impactando el comercio local (Oyanedel et al., 2022).

Una serie de parámetros químicos, físicos y biológicos se han estudiado a lo largo de los años en países como Chile, Nueva Zelanda, Polonia y Estados Unidos, encontrándose similitudes en la calidad de agua de estos países, pero aun así no es claro qué parámetros afectan o impulsan el crecimiento y supervivencia del Didymo, a excepción del fósforo, cuya presencia en bajas concentraciones se ha impuesto como una variable relevante en la proliferación masiva de esta microalga (Díaz et al., 2017; Oyanedel et al., 2022). En consecuencia, se han realizado múltiples investigaciones, y algunas han incluido análisis estadísticos y modelos de clasificación, no obteniendo resultados confiables. En el estudio de Hix & Murdock, 2019, se realizó un modelo de clasificación *Random Forest*, con una tasa de error del 18% en promedio para muestras con presencia de Didymo y un error del 100% para muestras con ausencia de Didymo (Hix & Murdock, 2019). Esto es un claro ejemplo de la complejidad al momento de analizar y modelar los datos que se obtienen como resultado de muestreos realizados en terreno.

Actualmente no se tienen claro los rangos de los parámetros que interactúan en la supervivencia de Didymo, y con los antecedentes revisados, es que se propone utilizar técnicas de inteligencia artificial para dar un nuevo enfoque a las investigaciones actuales en Chile y realizar proyecciones para futuros monitoreos con los datos que se han recopilado desde el año 2016 hasta la fecha por el Instituto de Fomento Pesquero, dentro del Proyecto: Convenio Desempeño 2021-2022: Monitoreo de la especie plaga *Didymosphenia geminata* en cuerpos de agua de la zona centro, sur y austral de Chile, Etapa VI, 2021-2022 (Oyanedel et al., 2022).

1.1 Justificación

En Chile, la Resolución Exenta 1854/2022 de la Subsecretaría de Pesca y Acuicultura del Ministerio de Economía, Fomento y Turismo, ha declarado 151 subcuencas hidrográficas con plaga y 29 subcuencas hidrográficas con riesgo plaga de Didymo en cuerpos de agua terrestre, pertenecientes a la zona centro y sur del país (subcuenca se define como un curso de agua que desemboca en un cuerpo de agua de mayor tamaño), y año tras año, el número de subcuencas afectadas ha ido incrementando (R.Ex.N°1854-2022: Declara Área de Plaga y de Riesgo de Plaga Que Indica En Cuerpos de Agua Que Señala En Materia de Acuicultura, 2022).

Consultoras medio ambientales, institutos de investigación, universidades, entre otros, han realizado múltiples monitoreos y han registrado datos *in situ* en la lucha contra el Didymo en distintas subcuencas, creando manuales de monitoreo y prevención, sin embargo, se ha planteado una colaboración entre los autores del presente documento y el Instituto de Fomento Pesquero de Puerto Montt para extraer información y conocimientos de los datos almacenados.

1.2 Planteamiento del trabajo

Para extraer el máximo conocimiento de estos datos, es que se ha propuesto al Instituto de Fomento Pesquero de Puerto Montt, división acuicultura, el estudio de los parámetros monitoreados a través de herramientas de inteligencia artificial, haciendo una nueva similitud entre ellos a través de *clustering*, para luego realizar un modelo de clasificación y así, en futuros monitoreos, conocer el estado y la presencia de Didymo de acuerdo con los parámetros químicos medidos, y finalmente un modelo de regresión lineal para predecir el avance de la plaga en años futuros.

1.3 Estructura de la memoria

La presente memoria cuenta con 5 capítulos. Comenzando con la introducción, en donde se entregan datos generales sobre el problema y se expone brevemente la propuesta a realizar.

El segundo capítulo corresponde al estado del arte en donde se hará una revisión profunda sobre Didymo, incluyendo la historia, hábitat, morfología, propagación, impacto, parámetros de crecimiento, registro de parámetros, geografía, beneficios potenciales, entre otros. Adicionalmente, se realizará una revisión exhaustiva sobre análisis exploratorio de datos, algoritmos de clasificación, *ensembled learning* y redes neuronales.

El tercer capítulo expondrá los objetivos a realizar, detallando el objetivo general y los objetivos específicos y la metodología de trabajo. La metodología explicará la limpieza de datos, utilización de herramientas como *clustering*, *ensembled learning* y redes neuronales perceptrón multicapa y convolucionales para grafos.

El cuarto capítulo presentará el desarrollo de la solución propuesta y los resultados obtenidos, la comparación entre ellos y comparación con bibliografía.

Finalmente, en el quinto capítulo se detallarán las conclusiones obtenidas, el cumplimiento de objetivos general y específicos y propuestas para líneas de trabajo futuro.

2. Contexto y estado del arte

El contexto y estado del arte del presente trabajo de fin de máster está dividido en dos partes. En la primera parte se hará una revisión exhaustiva sobre Didymo, evidenciando y contextualizando el problema a estudiar y justificando la información con estudios de numerosos autores.

La segunda parte explicará diversas herramientas de análisis de datos e inteligencia artificial que respaldarán la solución propuesta al problema.

2.1. Didymosphenia geminata

2.1.1. ¿Qué es *Didymospheniea geminata*?

Didymosphenia geminata (Didymo) es una especie de alga diatomácea, comúnmente conocida como "moco de roca" o "diatomea de piedra" (Figura *1*), que se encuentra en diversos sistemas fluviales en todo el mundo (Lamaro Anabel A. et al., 2019). Esta especie se caracteriza por su capacidad para formar colonias densas y notables que cubren el lecho de ríos y arroyos.

Las colonias de Didymo son conglomerados macroscópicos pueden tener una apariencia similar a un material mucoso o gelatinoso, y su color varía generalmente desde tonos amarillos hasta marrones (Department of Conservation, 2011; Díaz et al., 2011).

Figura 1: Aspecto de *Didymosphenia geminata.* Fuente: (Diaz et al., 2016)

2.1.2. Historia

Didymo históricamente es originaria de los ríos y lagos alpinos y boreales de Norteamérica y Europa (Iturrieta, 2016), fue descubierta en 1817 en las Islas Faroe (Blanco & Ector, 2009) y fue reportada por primera en América del Norte a mediados de 1850 (Bravo et al., 2019).

Los registros más antiguos de colonias o floraciones de Didymo masivas se remontan a más de un siglo en los ríos de Escandinavia, mientras que, en los Estados Unidos y Canadá, los registros más antiguos de tales floraciones datan de solo dos décadas atrás, específicamente en 1988, en el río Herber, Isla de Vancouver, Columbia Británica. En esta primera instancia, se evidenció la presencia de cúmulos viscosos pegados en rocas, un año después, varios kilómetros estaban cubiertos de esta alga (Bothwell et al., 2014). Desde entonces, científicos del hemisferio norte han observado y documentado la presencia de extensas floraciones en diversas ubicaciones, que incluyen Canadá, Estados Unidos, Islandia, Polonia, Italia y España. Es importante destacar que, en aquel entonces, estas manifestaciones no se consideraron invasiones, ya que Didymo ya estaba presente en esos países. Sin embargo, en la última década, esta especie ha sido reconocida frecuentemente como una invasora altamente agresiva en ríos oligotróficos de todo el mundo (Lamaro Anabel A. et al., 2019).

En el hemisferio sur, Didymo se detectó por primera vez en Nueva Zelanda en 2004 en el río Waiau, por un grupo de investigadores del Instituto Nacional de Investigación del Agua y la Atmósfera o NIWA (por sus siglas en inglés), detectaron por primera vez un brote masivo de Didymo (Kilroy, 2004).

En Chile, el primer reporte de Didymo fue en el año 1962 en el río Cisnes y río Sarmiento, pero en 2010 en el río Futaleufú se registra por primera vez de forma masiva (Díaz et al., 2016; Esse et al., 2018), luego, se ha extendido por toda la zona centro y sur del país, desde la región de Maule hasta la región de Magallanes y la Antártica Chilena, colonizando altas latitudes, como los sistemas fluviales en la región de la Patagonia del sur de Chile y Argentina y habitando principalmente sustratos rocosos, siendo declarada plaga en más de 70 subcuencas en la Resolución Exenta 719/2021 de la Subsecretaría de Pesca y Acuicultura (AMAKAIK Consultoría Ambiental, 2014;SUBPESCA, 2021).

2.1.3. Morfología

Didymo es una microalga de la familia de las diatomeas de agua dulce que genera proliferaciones mucilaginosas en aguas oligotróficas (Salvo & Oyanedel, 2019).

Para entender la definición anterior, debemos entender que son las diatomeas en primer lugar. Las diatomeas son algas microscópicas unicelulares, rodeadas de una pared de sílice (conocida como frústula) y compuesta por dos valvas (forma de plato plano) y bandas conectoras. Miden entre 5 y 200 µm y pueden llegar a medir hasta 1 mm. Se encuentran en distintas formas, como asociadas a filamentos, formando cadenas celulares, dentro de tubos mucilaginosos, viviendo libremente en la columna de agua o adheridas a cualquier sustrato (cuerpos sólidos). Son un componente importante del fitoplancton, bentos (organismos del fondo del ecosistema acuático) y las comunidades de algas de agua dulce y marina (Cameron, 2013; Sabater, 2009).

Como se observa en la Figura 2, Didymo es de forma alargada y presenta tallos mucilaginosos formados por mucopolisacaridos sulfatados que se secretan a través de una fila de poros ubicados en el polo basal de la pared celular, siendo la única diatomea de agua dulce con esa característica. Puede medir a lo largo entre 125 a 140 µm y 35 a 45 µm a lo ancho. En la naturaleza se puede encontrar adherida a superficies sólidas, de preferencia rugosas, de forma individual sin tallos, o en forma de colonia, formando mantos gelatinosos, extensos y gruesos, con un grosor de hasta 20 cm y de color gris a verde marrón. A pesar de ser visualmente viscosa, al tacto tiene una textura esponjosa (Bothwell et al., 2009; Bravo et al., 2019; Burkholder, 2009; Sheath & Wehr, 2015; Sterrenburg et al., 2007; Whitton et al., 2009; Y. Wu, 2017) y su crecimiento se ve afectado por factores como la temperatura, luminosidad, movimientos de las aguas, pH, conductividad eléctrica y la concentración de distintos nutrientes como fósforo, nitrógeno y calcio (Watson et al., 2015).

Figura 2: Morfología de Didymo. Arriba: valva de una célula vida. Centro: frústula vacía. Abajo: Células unidas a tallas mucilaginosos. Fuente: (Kilroy, 2004).

2.1.4. Hábitat

Todos los ríos mencionados en la historia de Didymo tienen en común 2 elementos, aguas frías y baja concentración de nutrientes, por lo que esta microalga crece en agua dulce (lagos, ríos y arroyos), en condiciones de pH neutro o levemente alcalino, temperatura de las aguas frías a cálidas, pobre de nutrientes y limitada en fosfato, caudal bajo y flujo constante y con alta luminosidad (Añón Suárez D, Albariño R, 2020; Gouvernement du Québec, 2008; SUBPESCA, 2010).

La temperatura ideal para su proliferación es entre 6 y 20 °C, con alta luminosidad y exposición al sol y en aguas de baja profundidad, entre 10 cm a 1.5-2.0 metros (Bravo et al., 2019; Kilroy, 2004; Larned et al., 2007).

Con respecto al caudal, este debe ser bajo, pero en constante movimiento, así permite que Didymo se adhiera a sustratos como rocas, plantas y los bordes de lagos, creando redes de gran extensión (en el río Futaleufú alcanzó 5 kilómetros), y espesores de 20 centímetros (Segura, 2011). Algunos autores han reportado que cuando Didymo forma están grandes extensiones, un alto caudal del agua beneficia la proliferación, ya que la adhesión aumenta y crea una biomasa compacta (Kilroy, 2004). En la Figura *3* se observa floración masiva de Didymo.

Didymo crece en ambientes oligotróficos, la falta de nutrientes y principalmente la baja concentración de fósforo estimulan el crecimiento de los pedúnculos de fijación (Salvo & Oyanedel, 2019).

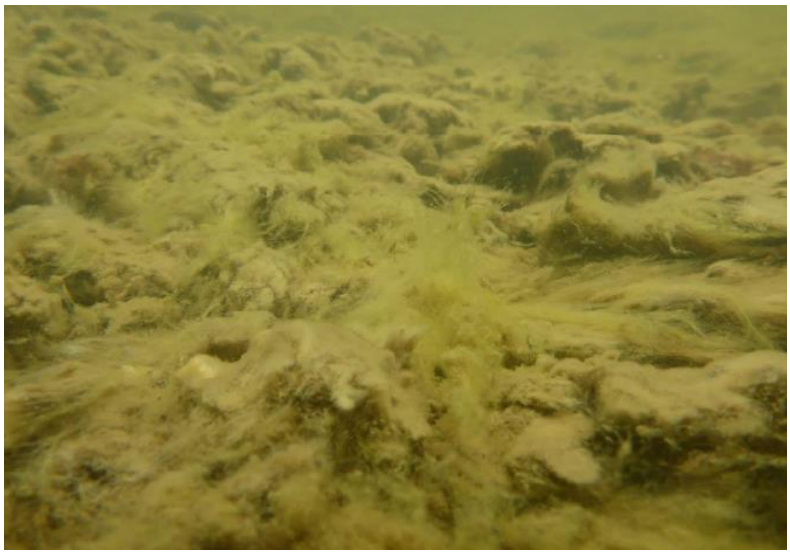


Figura 3: Floración masiva de Didymo. Fuente: (Oyanedel et al., 2022)

2.1.5. Propagación de Didymo

Didymo se ha propagado por todo el mundo en un corto periodo de tiempo, lo cual se atribuye principalmente al turismo, actividades como "*kayacking*", "*rafting*" y la pesca deportiva son los principales responsables de la introducción de esta especie. A través de equipos de pesca, botas, kayaks, vehículos, entre otros, sin la limpieza adecuada, el humano a transportado estas células vivas desde sitios contaminados hacia zonas libres de Didymo (Betancurt et al., 2016).

Algunos autores han mencionado el transporte de Didymo a través de aves y animales. Las aves pueden transportar células vivas adheridas en patas y plumas, esparciendo esta alga en los procesos de migración (dispersión local más que global), y algunos mamíferos pueden llevar esta alga en el pelaje. También, Didymo podría sobrevivir en el tracto digestivo de estos animales. El viento se considera otro factor en la propagación de Didymo en cortas distancias (Kilroy, 2004; SUBPESCA, 2010).

2.1.6. Impacto

Didymo puede alterar completamente el ecosistema de un sistema fluvial. El Didymo puede generar extensas colonias al adherirse a las rocas o plantas y cubrir el fondo del curso de agua (Department of Conservation, 2011).

Los sistemas fluviales hemisferio sur, en especial de la zona patagónica de Argentina y Chile, son calificados como vulnerables a nuevas introducciones e invasiones de Didymo (Lamaro Anabel A. et al., 2019), lo que conlleva una variedad de impactos, los cuales detallamos a continuación:

- **Ecológico**

  En el primer caso, como consecuencia al crecimiento de las colonias del alga, se modifica el aspecto visual de los sitios y éstos tienden a percibir como contaminados". Sin embargo, el mayor impacto físico para el ambiente ocurre por la marcada retención de sedimentos que producen el Didymo y la alteración de la hidrodinámica del sistema fluvial. En el caso de los procesos biogeoquímicos, el efecto del crecimiento masivo del alga es que producen cambios en el pH lo que modifica las condiciones químicas de los sitios afectados, influyendo en la disponibilidad de nutrientes, originando el desplazamiento de las comunidades algales nativas. A nivel de las cadenas tróficas, existe evidencia de afectación en la población de los grupos de macroinvertebrados (como insectos, crustáceos, o anélidos.(Baffico & Beamud, 2017). Esta alteración en el ecosistema afecta la abundancia de peces de los ríos (Tapia, 2012).

- **Socioeconómico**

  Las densas colonias de Didymo tienen un impacto significativo en la utilización de los ríos, provocando una reducción drástica en su valor recreativo y estético. De hecho,

estas acumulaciones de diatomeas asemejan visualmente los vertidos industriales, lo que disminuye la calidad visual de los cuerpos de agua. Además, la necesidad de limpiar embarcaciones y equipos de pesca implica costos económicos considerables. Adicionalmente, se han observado problemas de obstrucción en canales, centrales hidroeléctricas, plantas de tratamiento de aguas residuales y plantas potabilizadoras debido a la acumulación de las masas mucosas de Didymo. Esto requiere limpiar estas estructuras, lo que conlleva gastos económicos adicionales (Capdevila-Argüelles et al., 2011).

- **Sanitario**

  A pesar de no ser una especie intrínsecamente tóxica, las personas que se bañan en áreas donde Didymo se encuentra en grandes concentraciones pueden experimentar irritación ocular debido a la sílice presente en las paredes celulares de esta especie (Capdevila-Argüelles et al., 2011).

2.1.7. Parámetros físicos y químicos

Dentro del hábitat de Didymo se destacan los ambientes oligotróficos. Estos son ambientes con baja concentración de nutrientes, poca producción de fitoplancton y predominan durante el invierno (Labib et al., 2023).
Múltiples parámetros físicos y químicos intervienen en la proliferación y crecimiento de Didymo, dentro de los cuales encontramos:

- Temperatura: es la propiedad de un cuerpo de transmitir calor desde o hacia otros cuerpos. El Didymo se ha encontrado en aguas con temperatura ente 0.1°C y 27°C (Beeby, 2012; Hanna Instruments, 2023).

- Luminosidad: ambientes con luz directa durante el día, sin sombra. La intensidad de la luz incrementa la división celular (Hix & Murdock, 2019). Este parámetro no está cuantificado.

- Velocidad del agua: Didymo se desarrolla en aguas con baja velocidad, rango entre de 0 m/s a 4.5 m/s. Si la velocidad es alta, no se adhiere al sustrato, evitando la proliferación (Iturrieta, 2016).

- Turbidez: en agua, es una propiedad óptica de la luz que se dispersa y absorbe, en vez de transmitirse. Al encontrar partículas sólidas suspendidas en el agua, la luz se dispersa. A mayor turbidez, mayor es la cantidad de partículas suspendidas. En el crecimiento de Didymo, la turbidez debe ser baja, con mediciones inferiores a 5 NTU, sin embargo, en ríos del sur de Chile se han reportado zonas libres de Didymo con mediciones sobre los 10 NTU (Hanna Instruments, 2023; Jellyman et al., 2006; Salas et al., 2014).

- pH: es la medición de la concentración del ion hidrógeno [H^+] en agua o sólidos. Va desde una escala de 0 a 14, en donde 7 es neutro, menor a 7 es ácido y mayor a 7 se considera básico o alcalino. El pH óptimo para el crecimiento de diatomeas, y Didymo, es entre 6.4 y 9.0. También, juega un papel importante en el desarrollo de frústulas, ya que el pH ácido incrementa la solubilidad del silicato y afecta negativamente la formación de estas (Beeby, 2012; Hanna Instruments, 2023; Hix & Murdock, 2019).

- Conductividad eléctrica: es la medición de cuan bien un material conduce cargas eléctricas y se mide en µS/cm. El rango es entre 20 y 120 µS/cm, sin embargo, no es un parámetro determinante en la proliferación (Hanna Instruments, 2023; Tapia, 2012).

- Salinidad: es la medición de la cantidad de sales disueltas en un sistema. Se mide a partir de la conductividad eléctrica (Hanna Instruments, 2023).

- Oxígeno disuelto: es la medición de cuanto oxígeno está disuelto es una sustancia líquida y se mide en mg/L o en porcentaje de saturación (%). Se ve afectado por la temperatura, la salinidad y la presión. El rango óptimo para el crecimiento de Didymo es de 92% a 140% (Hanna Instruments, 2023; Tapia, 2012).

- Calcio: ayuda a las plantas a desarrollar e incrementar resistencia en los tejidos vegetales y tallos y ayuda en la formación de adhesivos para adherirse a los sustratos. El rango de calcio va entre 3.15 mg/L y 10.9 mg/L (Sutherland et al., 2007), sin embargo, se ha encontrado evidencia de rangos entre 21.2 mg/L y 85 mg/L e incluso entre 1.1 mg/L y 170.2 mg/L (Hanna Instruments, 2023; Hix & Murdock, 2019; Kawecka & Sanecki, 2003).

- Magnesio: es un mineral que ayuda a la producción de clorofila, pigmento de color verde que absorbe luz y provee de energía a las plantas. También, aumenta la concentración de vitaminas y ayuda en la absorción de fosforo, y al igual que el calcio, ayuda en la formación de adhesivos para adherirse a los sustratos. Se ha evidenciado que el rango en el que se encuentra es entre 4.4 mg/L y 19.5 mg/L (Hanna Instruments, 2023; Hix & Murdock, 2019; Kawecka & Sanecki, 2003).

- Dureza carbonatada: es la cantidad de sales cálcicas disueltas en agua. Puede encontrarse en un rango entre 4.5 mg/L y 12.6 m/L (Hanna Instruments, 2023; Kawecka & Sanecki, 2003)

- Alcalinidad: es la capacidad del agua de mantener el pH estable, cuando este valor es bajo, es difícil de mantener el pH estable, entre más alto el valor, más difícil de que se produzcan cambios. El rango de alcalinidad va entre 11.8 mg/L a 29.7 mg/L como $CaCO_3$/L (Hanna Instruments, 2023; Sutherland et al., 2007).

- Fosfato: es esencial en el crecimiento de raíces, tallos flores y semillas de plantas. En niveles altos puede causar eutrofización (contaminación, película en la superficie del agua, por lo general de color verde, producto del exceso de nutrientes, como nitrógeno y fosfato). El rango es entre 0.001 mg/L y 1.2 mg/L (Hanna Instruments, 2023).

- Fosforo total: es un mineral esencial para el crecimiento de plantas y animales, sin embargo, si está disuelto en agua en alta concentración, puede causar crecimiento excesivo de algas y microorganismos. Para Didymo, en bajas concentraciones promueve el crecimiento de los tallos y en altas concentraciones, se observa un crecimiento retardado de estos. El rango en el que se encuentra es entre 0.0004 mg/L y 2.1 mg/L (Hanna Instruments, 2023; Hix & Murdock, 2019; Sundareshwar et al., 2011). Este parámetro podría ser determinante en la presencia o ausencia de Didymo, en estudios han reportado que el fosforo en altas concentraciones (en promedio 75 mg/L) evita la proliferación de Didymo (Bravo et al., 2019).

- Hierro: los tallos de Didymo absorben primero hierro y luego fosforo, luego las enzimas y bacterias interactúan con el hierro para que se capture mayor cantidad de fosforo. Puede encontrarse en un rango entre 0.006 mg/L y 3.513 mg/L (Sundareshwar et al., 2011).

- Nitrato: se forma por la descomposición de material orgánico y en rango altos puede causar eutrofización. Puede estar en un rango entre 0.7 mg/L y 14 mg/L, sin embargo, se ha encontrado evidencia de un rango entre 0.08 mg/L y 2.9 mg/L o entre 0.003 mg/L y 79.3 mg/L (Hanna Instruments, 2023; Kawecka & Sanecki, 2003; Tapia, 2012).

- Nitrito: es un producto intermedio entre el ciclo del nitrógeno y es producto de la oxidación de amoniaco. Puede encontrarse en un rango de 0.001 mg/L y 0.35 mg/L (Hanna Instruments, 2023).

- Nitrógeno Total: es la suma de todas las formas de nitrógeno (nitrato, nitrito, nitrógeno total Kjeldahl y amoniaco). En concentraciones altas, puede eliminar el oxígeno disuelto en agua, impactando negativamente la vida de algas y animales acuáticos. El rango óptimo de este parámetro para el crecimiento de Didymo es entre 0.044 mg/L y 0.158 mg/L (Hanna Instruments, 2023; Hix & Murdock, 2019; Tapia, 2012).

- Silicato: como se mencionó anteriormente, la pared celular de Didymo está compuesta por silicato, pero también se encuentra disuelto por en agua. Se trabaja en con el rango para agua entre 0.2 mg/L y 83.8 mg/L (Salvo & Oyanedel, 2019).

- Cloruro: es importante para el equilibrio de iones osmótico (transporte de agua) y el intercambio osmótico a través de la membrana celular de la célula. Pueden encontrarse en un rango entre 0.8 mg/L y 2.2 mg/L, sin embargo, dependiendo de la ubicación geográfica de las aguas, pueden encontrarse entre 1 mg/L y 3 mg/L (Hix & Murdock, 2019; Tapia, 2012).

- Sulfato: tiene relación con las propiedades biológicas de las algas y plantas, como antioxidantes y anticoagulantes, y está regulado por proteínas de la membrana celular. Algunos estudios indican que podría tener relación con el alto contenido de sulfuro en Didymo. Puede encontrarse en un rango entre 0.74 mg/L y 2.5 mg/L, sin embargo, no es determinante en la proliferación de esta especie (Hix & Murdock, 2019; Salas et al., 2014).

2.1.8. Parámetros que afectan la supervivencia de Didymo

Dentro de las investigaciones realizadas sobre Didymo, se han estudiado parámetros que ayudarían a combatir esta alga

- Cobre: es un metal que, disuelto en agua, es tóxico para algas y peces, y se ha demostrado que, a una menor temperatura de las aguas, este metal se vuelve aún más toxico, afectando la supervivencia de Didymo. Se puede encontrar en compuestos químicos para el control de algas y hongos, como cobre quelado (Beeby, 2012; Jellyman et al., 2006)

- Zinc: el zinc es un nutriente y forma parte de compuestos químicos utilizados en el control de algas en piscinas (sulfato de zinc). En concentración alta, la supervivencia de Didymo disminuye, pero con menor efectividad comparada con el cobre (Beeby, 2012; Jellyman et al., 2006).

- Cloro: es un desinfectante y blanqueador y afecta a la supervivencia de bacterias y algas. El compuesto químico más utilizado para desinfectar agua potable, piscinas y torres de enfriamiento es el hipoclorito de sodio (Jellyman et al., 2006).

2.1.9. Registro de parámetros y muestreo

El muestreo de Didymo cuenta con 3 fases: elección del sitio e inspección visual, muestreo físico y químico y muestreo biológico.

La elección del sitio e inspección visual contempla actividades como la elección del tramo de muestreo, georreferenciación y descripción visual. La elección del sitio tiene relación con las actividades que se desarrollan en el lugar*, kayaking*, pesca deportiva, entre otras, y se eligen puntos que ya cuentan con presencia de Didymo, lugares que anteriormente fueron muestreados, pero no mostraron trazas de Didymo y posibles lugares en donde podría desarrollarse la plaga (Diaz et al., 2016).
Parámetros como temperatura, pH, conductividad eléctrica, oxígeno disuelto y turbidez, se miden directamente en el sitio con instrumentos especializados, rápidos y precisos. Los datos se registran manualmente y luego son transferidos a una tabla para su posterior almacenamiento.

Parámetros como nitrato, nitrito, nitrógeno, fosfato, fósforo, silicato, calcio y hierro, son llevados en contenedores específicos para cada análisis, preservando cada muestra a una temperatura de 4°C en su traslado, hasta arribar a un laboratorio acreditado. Los datos se obtienen en un documento certificado y posteriormente son agregados a la tabla de los parámetros en terreno manualmente (Díaz et al., 2017).

Para el muestreo biológico, se realiza en primer lugar una inspección visual con el fin de detectar la presencia de Didymo, existiendo una clasificación para el manto de la mucosidad, el cual se detalla en la Tabla *1*. Luego, fitoplancton y especies bentónicas, son consideradas al momento de realzar en análisis taxonómico en un laboratorio (Diaz et al., 2016).

Tabla 1: Clasificación de presencia/ausencia de Didymo según es espesor del manto de la mucosidad a través de inspección visual.

Estado	**Espesor**	**Cobertura**	**Definición**
Ausente	Espesor = 0	Ausente	No se observa mucosidad
Inicial	Espesor < 0,2 cm	Cobertura < 20%	Crecimiento inicial
Mediana	0,2 cm ≤ espesor < 1 cm	20% ≤ cobertura < 50%	Mucosidad parchosa
Alta	1 cm ≤ espesor < 2 cm	50% ≤ cobertura < 80%	Zonas compactadas y pequeñas agregaciones
Muy alta	Espesor ≥ 2 cm	Cobertura ≥ 80%	Manto compacto

Fuente: (Manual para el monitoreo e identificación de la microalga bentónica Didymosphenia geminata, 2016)

2.1.10. Geografía

Didymo ha sido declarada plaga y riesgo de plaga en numerosas cuencas y subcuencas a lo largo de la zona centro sur de Chile, extendiéndose desde la Región del Maule hasta la Región de Magallanes. Una cuenca hidrográfica es superficie terrestre en donde el agua es captada y almacenada, y desagua en ríos, lagos, lagunas, humedales, estuarios, pantanos, entre otros. Están formadas por subcuencas, que son lugares de almacenaje de agua de menor tamaño que la cuenca y son los afluentes de la cuenca (Diaz et al., 2016). Las cuencas con presencia de Didymo (plaga o riesgo) establecidas por la Subsecretaría de Pesca y Agricultura del Gobierno de Chile se detallan en la Tabla *2*:

Tabla 2: Detalle de las cuencas declaradas con plaga Didymo.

	Región	**Cuenca**
1	Maule	Río Maule
2	Ñuble	Río Itata
3	Biobío	Río Biobío
4		Río Imperial
5	La Araucanía	Río Toltén
6		Río Valdivía
7		Río Bueno
8	Los Ríos	Cuencas e Islas entre Río Bueno y Río Puelo
9		Río Puelo
10	Los Lagos	Coteras entre Río Puelo y Río Yelcho
11		Río Yelcho
12		Río Palena y Costeras Limite Décima Región
13		Costeras e Islas entre Río Palena y Río Aysen
14	Aysen del General Carlos Ibañez del Campo	Río Aysen
15		Costeras e Islas entre Río Aysen y Río Baker y Canal General Martínez
16		Río Baker
17		Costeras e Islas entre Río Baker y Río Pascua
18		Río Pascua
19	Magallanes y de la Antártica Chile	Costeras entre Seno Andrew y Río Hollemberg e islas al oriente
20		Tierra del Fuego

Fuente: (R.Ex.N°1854-2022: Declara Área de Plaga y de Riesgo de Plaga Que Indica En Cuerpos de Agua Que Señala En Materia de Acuicultura, 2022)

2.1.11. Efectos en las Cuencas Hidrográficas de Chile

La riqueza de macroinvertebrados en diferentes regiones de Chile en relación con la presencia de Didymo, que ha demostrado tener un impacto significativo en los hábitats acuáticos de Chile (Díaz et al., 2017).

En el estudio de la comunidad de macroinvertebrados bentónicos, se logró identificar la mayoría de ellos hasta el nivel taxonómico de familia. No obstante, algunos taxa específicos, como Hirudinea, Oligochaeta, Nematomorpha, Collembola, Acari y Turbellaria, fueron descritos solo hasta ciertos niveles taxonómicos, como se detalla en la Tabla *3* del estudio. Para poder evaluarlos en conjunto con los demás macroinvertebrados, se agruparon en categorías de familias indeterminadas (Díaz et al., 2017).

Se registraron un total de 49 taxa de macroinvertebrados en las cinco zonas analizadas. En la zona del Biobío, fueron registrados 24 taxa, siendo los más abundantes Leptophlebiidae (36.54%), Hydrobiidae (21.80%) y Baetidae (10.13%); en la Araucanía 40 taxa, siendo lo más abundantes Chironomidae (24.02%), Leptophlebiidae (19.59%) y Baetidae (12.48%); en Los Ríos fueron registrados 44 taxa, siendo los más abundantes Chironomidae (22.65%), Leptophlebiidae (21.79%) y Hydropsychidae (16.25%); en Los Lagos fueron registrados 32 taxa, siendo los más abundantes Chironomidae (25.19%), Leptophlebiidae (20.39%) y Hydropsychidae (9.19%); y en la región de Aysén fueron registrados 37 taxa, siendo los más abundantes Chironomidae (35.35%), Leptophlebiidae (15.01%) y Leptophlebiidae (13.87%), (Tabla *4*) (Díaz et al., 2017).

Tabla 3: Macroinvertebrados bentónicos que no fueron identificados hasta nivel taxonómico de familia.

Phylum	**Clase**	**Orden**	**Familia**
Annelida	Hirudinea	Indeterminado	Familia 1 indeterminada
Annelida	Oligochaeta	Indeterminado	Familia 2 Indeterminada
Nematomorpha	Indeterminado	Indeterminado	Familia 3 indeterminada
Platyhelminthe	Turbellaria	Indeterminado	Familia 4 indeterminada
Arthropoda	Arachnoidea	Acari	Familia 5 Indeterminada
Arthropoda	Entognatha	Collembola	Familia 6 Indeterminada

Fuente: (Díaz et al., 2017).

Tabla 4: Macroinvertebrados bentónicos identificados hasta nivel taxonómico de familia.

Taxa/Regiones	**Biobío**	**Araucanía**	**Los Ríos**	**Los Lagos**	**Aysén**
Acari	X	X	X	X	X
Aeglidae	X	X	X	X	X
Aeshnidae	X	X	X	X	X
Ameletopsidae		X	X	X	X
Ancylidae	X	X	X		

Taxa/Regiones	Biobío	Araucanía	Los Ríos	Los Lagos	Aysén
Athericidae	X	X	X	X	X
Austroperlidae	X	X	X	X	X
Baetidae	X	X	X	X	X
Blephariceridae	X	X	X		X
Ceratopogonidae		X	X	X	X
Chilinidae	X	X	X	X	X
Chironomidae	X	X	X	X	X
Collembola		X			
Corydalidae	X	X	X	X	
Diamphipnoidae		X	X		
Ecnomidae		X	X		X
Elmidae	X	X	X	X	X
Empididae	X	X	X	X	X
Eustheniidae			X		
Glossosomatidae	X	X	X	X	X
Gripopterygidae	X	X	X	X	X
Gyrinidae			X		
Hirudinea		X	X		X
Hyalellidae	X	X	X		X
Hydraenidae			X		
Hydrobiidae	X	X	X	X	X
Hydrobiosidae		X	X	X	X
Hydrophilidae			X		
Hydropsychidae	X	X	X	X	X
Hydroptilidae		X	X	X	X
Leptoceridae		X	X	X	X
Leptophlebiidae	X	X	X	X	X
Limnephilidae			X		X
Lymnaeidae				X	X
Muscidae			X		X
Nematomorpha		X	X	X	X
Notonemouridae		X	X	X	X
Oligochaeta	X	X	X	X	X
Oniscigastridae		X	X		

Taxa/Regiones	Biobío	Araucanía	Los Ríos	Los Lagos	Aysén
Perlidae		X	X	X	X
Physidae		X	X		
Polycentropodidae				X	X
Psephenidae	X	X	X	X	
Pyralidae			X		X
Sericostomatidae		X		X	X
Sialidae		X			X
Simuliidae	X	X	X	X	X
Tipulidae	X	X	X	X	X
Turbellaria	X	X	X	X	
49	24	40	44	32	37

Fuente: (Díaz et al., 2017).

En el análisis de la composición y abundancia de macroinvertebrados bentónicos en ríos con y sin la presencia de Didymo, se observaron diferencias estadísticamente significativas ($p<0.05$). Esto indica que la composición de macroinvertebrados y sus cantidades variaron notablemente en ríos donde Didymo estaba presente en comparación con aquellos donde no estaba presente. Asimismo, se obtuvieron resultados similares al comparar las diferentes regiones geográficas ($p<0.05$), lo que llevó al análisis separado de los datos según las regiones.

Estos hallazgos sugieren que la presencia de Didymo ejerce un impacto significativo en la comunidad de macroinvertebrados bentónicos y que estas diferencias también varían según la región geográfica, lo que puede estar relacionado con las condiciones ambientales y ecológicas específicas de cada área (Díaz et al., 2017).

En las estaciones donde se encontró la presencia de Didymo, se identificaron un total de 39 taxa de macroinvertebrados bentónicos. Los taxa más abundantes en estas estaciones fueron Chironomidae, representando el 33.19% de la abundancia total, seguidos por Oligochaeta con el 14.47%, y Hydropsychidae con el 11.48%. Por otro lado, en las estaciones donde no se encontró Didymo, se registraron un total de 49 taxa de macroinvertebrados bentónicos. Los taxa más abundantes en estas estaciones fueron Leptophlebiidae, que representaron el 19.54% de la abundancia total, seguidos por Chironomidae con el 26.35%, y Baetidae con el 9.57% (Díaz et al., 2017). Estos resultados sugieren diferencias significativas en la composición y abundancia de macroinvertebrados

bentónicos entre estaciones con y sin la presencia de Didymo, lo que indica un posible impacto de esta especie invasora en la comunidad de macroinvertebrados, como se indica en la Figura *4*.

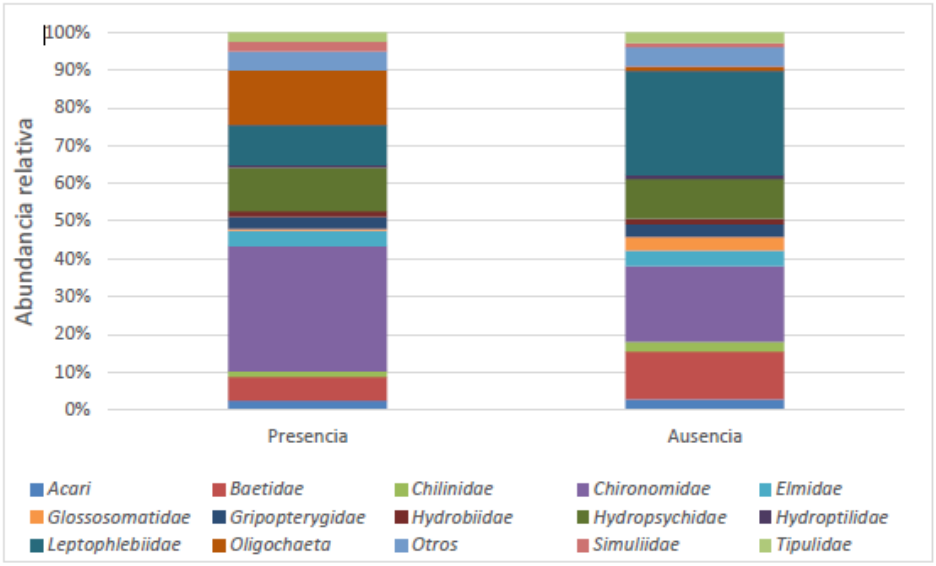


Figura 4:Representa la abundancia relativa de taxa de macroinvertebrados según presencia o ausencia de Didymo. Fuente: (Díaz et al., 2017).

2.1.12. Beneficios Potenciales

Anteriormente se indicó que en los últimos años, el Didymo, ha generado una creciente atención debido a su proliferación masiva en varios países, incluyendo Chile, por sus condiciones fisicoquímicos del agua, como la temperatura, la intensidad lumínica, las bajas concentraciones de fósforo total y fósforo disuelto, y caudales bajos. Condiciones que proporcionan un ambiente propicio para la expansión de Didymo en los cursos de agua (Figueroa et al., 2021).

Además, es importante destacar que Didymo se dispersa de manera pasiva, principalmente a través de la actividad humana en actividades recreativas, lo que ha llevado a un aumento en su área de distribución geográfica. Este fenómeno plantea preocupaciones ambientales debido a su impacto en los ecosistemas acuáticos (Figueroa et al., 2021).

Sin embargo, esta diatomea también ha despertado interés en la comunidad científica y en la industria debido a sus posibles aplicaciones. Se ha descubierto que Didymo posee actividad antioxidante y efectos tóxicos que pueden eliminar células de cáncer de Colon y leucemia humana. Además, estimulan una respuesta de activación en células del sistema

inmunológico, específicamente macrófagos, mediante la inducción de mediadores implicados en procesos inflamatorios como las citoquinas IL-6 y TNF-α, como se observa en la Figura *5*. Esto sugiere la posibilidad de convertir las proliferaciones colonias de Didymo que actualmente representan un problema ambiental, en una potencial investigación relacionada con el cáncer y la salud, abriendo nuevas perspectivas en la búsqueda de tratamientos y terapias relacionadas con enfermedades inflamatorias y ciertos tipos de cáncer, así en una fuente de polisacáridos con aplicaciones en la industria nutracéutica y cosmética (Figueroa et al., 2021).

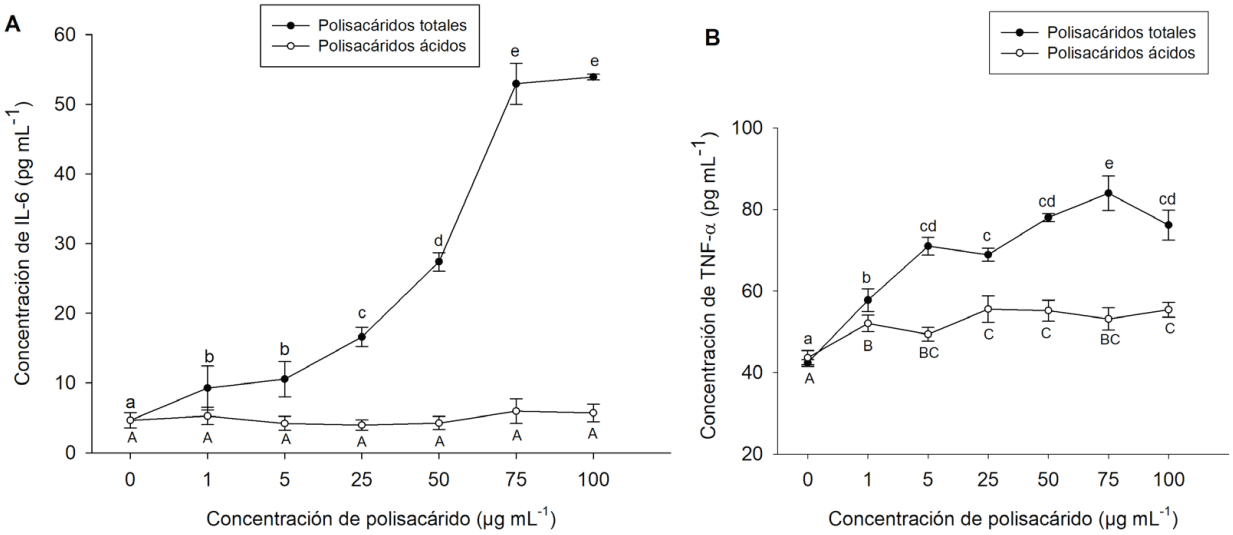


Figura 5: Gráfico representa el impacto de los polisacáridos totales y ácidos de Didymo en citoquinas inflamatorias mediadoras. A) Influencia en los niveles de Interleucina-6 (IL-6). B) Influencia en los niveles de Factor de Necrosis T. Fuente: (Figueroa et al., 2021)

En resumen, Didymo ha destacado por su proliferación en entornos acuáticos y su impacto en los ecosistemas, pero también presenta un potencial interesante en términos de aplicaciones en la medicina y la industria nutracéutica, lo que plantea nuevas perspectivas para su estudio y aprovechamiento (Figueroa et al., 2021).

2.2. Herramientas de análisis de datos e inteligencia artificial

2.2.1. Análisis exploratorio de datos

El análisis exploratorio de datos o EDA, por sus siglas en inglés (*exploratory data análisis*), y tal como lo indica su nombre, es una exploración en el set de datos para describir y estudiar la naturaleza de estos, identificar anomalías o datos fuera de rango, determinar relaciones entre las variables, verificar suposiciones, con el fin de entender y evaluar la calidad de los datos previo al procesamiento de estos (Komorowski et al., 2016; Ministerio de Asuntos Económicos y Transformación Digital, 2021; Seltman, 2018).

Hay cuatro tipos de EDA:

- Univariable no grafico: exploración de una sola variable o atributo o columna. Se estudian características como media aritmética, mediana, moda, rangos, varianza, desviación estándar, frecuencia, proporción entre otros. Estas mediciones se realizan dependiendo del tipo de dato (numérico, categórico, *booleano*) (Komorowski et al., 2016; Seltman, 2018).

- Multivariable no gráfico: estudio de la relación entre dos o más variables a través de técnicas como tabulación cruzada y correlación (Komorowski et al., 2016; Seltman, 2018).

- Univariable gráfico: estudio de una variable representada gráficamente en histogramas, diagramas de caja, diagramas de tallos y hojas o gráficos en dos dimensiones (Komorowski et al., 2016; Seltman, 2018).

- Multivariable grafico: estudio de dos o más variables representados en gráficos, como gráficos de dispersión, diagramas de caja lado-a-lado o regresiones lineales (Komorowski et al., 2016; Seltman, 2018).

Para EDA se describen los siguientes pasos a seguir:

- Análisis descriptivo de variables: estudio de las variables de forma gráfica y no gráfica, cálculo de medidas de tendencia central a través de estadística descriptiva, características del set de datos como el número de atributos e instancias (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

- Re-ajuste de los tipos de variables: verificación del tipo de variables a trabajar: numérica, categórica, *booleana*, fecha, entre otros (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

- Detección y tratamiento de datos ausentes y atípicos: el set de datos podría contar con datos ausentes, fuera de rango u otro tipo de anomalía, por lo que se deben estudiar en profundidad para decidir cómo abordarlos. Este tipo de dato se puede rellenar o reemplazar con medidas como media, mediana o moda, y no es recomendables eliminarlos del set de datos (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

2.2.2. Inteligencia artificial

Jhon McCarthy en 1955 definió la inteligencia artificial como la ciencia e ingeniería de hacer máquinas inteligentes. Los humanos programan máquinas para que realicen distintas tareas de forma inteligente, pero también, estas máquinas pueden aprender, cómo los seres humanos (Manning, 2020).
La inteligencia artificial engloba conceptos como *Machine Learning* y *Deep Learning*, siendo herramientas que ayudan a máquinas a imitar la inteligencia humana (Chollet, 2018).

2.2.3. Machine Learning

Se puede definir *Machine Learning* como el campo de estudio que le da a computadores la habilidad de aprender al programarlos (Géron, 2019), mejorando su percepción, conocimiento y pensamiento (Manning, 2020) y puede clasificarse de acuerdo con la cantidad y tipo de supervisión durante el entrenamiento de un modelo, entre ellas se destacan dos categorías: aprendizaje no supervisado y aprendizaje supervisado.

2.2.4. Aprendizaje no supervisado

El aprendizaje no supervisado son técnicas de *Machine Learning* que permiten descubrir patrones en un set de datos sin etiquetas determinadas (Raschka & Mirjalili, 2019). Existen varios tipos de aprendizaje no supervisado, como reducción de dimensionalidad, detección de anomalías y *clustering*, esta última se detallará en profundidad.

Clustering

Clustering es un método de aprendizaje no supervisado que consiste en dividir un set de datos en pequeños grupos de datos, de tal forma que los datos que pertenecen a un conjunto son muy similares entre sí, pero los datos de diferentes conjuntos no presentan similitud (Müller & Guido, 2017; Pajankar & Joshi, 2022). Esta técnica es utilizada en diferentes aplicaciones, como:

- Análisis de datos: en el análisis de nuevos sets de datos, *clustering* es una forma útil de descubrir instancias similares e incluso analizar los grupos por separado (Géron, 2019).

- Segmentación de clientes: agrupación de clientes por las compras que realizan, actividades en línea como visitas en sitios de internet y otros, lo que nos ayudará a entender quiénes son nuestros clientes y que necesitarán, dándole un nuevo enfoque a campañas de marketing para cada segmento (Géron, 2019).

- Detección de anomalías: las instancias con baja afinidad en los grupos se denomina anomalía. Esto es útil en la detección de defectos en la manufactura de productos o en la detección de fraude (Géron, 2019).

- Segmentación de imágenes: agrupando pixeles de acuerdo con el color y luego reemplazando cada píxel por el del color dominante, es posible reducir el número de diferentes colores de una imagen. Esto se utiliza en la detección de objetos y sistema de rastreo, lo que hace más fácil detectar el contorno de los objetos (Géron, 2019)

- Aprendizaje semi supervisado: si solo se tienen algunas etiquetas para el set de datos, se utiliza *clustering* para propagar las etiquetas de todas las instancias del

mismo grupo. Esto se hace con el fin de mejorar los modelos de aprendizaje supervisado posteriores (Géron, 2019).

Distintos algoritmos para técnicas de *clustering* se han desarrollado, y los métodos que utilizarán para el desarrollo de este trabajo, se detallarán a continuación:

Clustering k-means

Este es uno de los algoritmos más utilizado en la técnica de *clustering*, el que consiste en la división de un set de datos en pequeños grupos a través de un método basado en centroides. Trata de encontrar el centro de cada grupo que es representativo para cada conjunto a través del concepto de minimización de inercia (Ecuación 2.2): distancia media al cuadrado entre cada instancia y el centro más cercano, o distancia Euclidiana (Ecuación 2.1), y cada instancia es asignada a un único grupo (Pajankar & Joshi, 2022; Raschka & Mirjalili, 2019):

$$d(x,y)^2 = \sum_{j=0}^{m} \left(x_j - y_j\right)^2 \qquad \text{Ecuación 2.1}$$

En donde:
$d(x,y)^2$: cuadrado distancia Euclidiana
x e y : puntos para el cálculo de distancia
m : dimensión del espacio

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left|\left| x_i - \mu_j \right|\right|^2 \right) \qquad \text{Ecuación 2.2}$$

En donde:
μ_j: media, también llamada centroide
C : *cluster*
x_i: instancia
n : número de instancias

El centroide μ_j es calculado a través de la formula XXX:

$$\mu_j = \frac{1}{|N(x)|} \sum_{x_j \in N(x)} x_j \qquad \text{Ecuación 2.3}$$

En donde:

$N(x)$: número de instancias pertenecientes al grupo

x_j : instancia perteneciente al grupo j

El algoritmo de *k-means* se describe como:

1. Aleatoriamente se escogen *k* centroides los que serán los centros iniciales
2. Las instancias se asignan al centroide más cercano calculando la distancia más cercana (mínima inercia)
3. Se calcula un nuevo centroide para cada grupo (Ecuación 2.3)
4. Se repiten los pasos 2 y 3 hasta que la asignación de cada instancia no cambia o hasta que se alcance la cantidad máxima de iteraciones definidas por el usuario

Eligiendo el número de *clusters*

El algoritmo *k-means* espera que se ingrese un número de *clusters* y a partir de ese número realiza la división del set de datos. Este número de *clusters* k se determina a través del método del codo, calculando el error de la inercia, la cual representa cuan coherente son los grupos internamente. La inercia disminuye a medida que se incrementa número de *clusters* (Pajankar & Joshi, 2022; Raschka & Mirjalili, 2019).

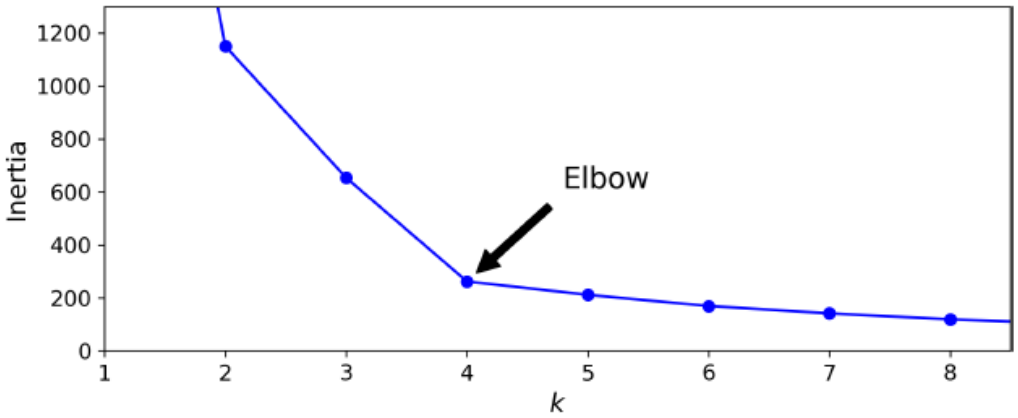


Figura 6: visualización del método del codo para la elección del número de *clusters*. Fuente: (Géron, 2019)

Como se observa en la Figura *6*, la inercia disminuye drásticamente hasta k=4, y luego disminuye lentamente. Este punto de inflexión corresponde al número óptimo de clusters a utilizar, si se consideran menos grupos los grupos tendrían poca afinidad, mientras que si

elegimos más grupos solo tendremos subdivisiones de grupos ya optimizados (Géron, 2019).

Calidad de un *cluster*

Una vez elegido el número de *clusters*, es importante determinar la calidad de estos. Para ellos se utiliza el coeficiente *Silhoutte*. Este coeficiente es un indicador que indica que tan alta es la similitud de los grupos, puede tomar valores entre -1 y 1. Si el coeficiente está más cercano a 1, indica que las instancias se clasificaron correctamente y están muy cerca del centroide, para valores cercanos a 0, las instancias están cerca de los bordes y para valores cercanos a -1, podría ocurrir que las instancias de clasificaron en un *cluster* erróneo (Géron, 2019; Raschka & Mirjalili, 2019).

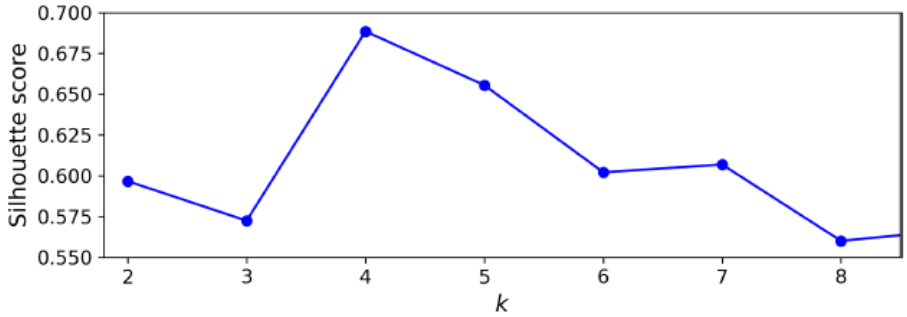


Figura 7: comparación de coeficiente *silhoutte* para distinto número de *clusters*. Fuente: (Géron, 2019)

Continuando con el ejemplo de la figura 6, al calcular el coeficiente *Silhouette*, se observa en la Figura *7* que k=4 sigue siendo el óptimo, con el valor del coeficiente por sobre los otros *cluster* y más cercano a 1.

El cálculo del coeficiente *silhouette* se realiza de la siguiente forma:

1. Calcular la cohesión del *cluster* $a^{(i)}$, como el promedio de la distancia entre un punto en específico y todos los otros puntos del mismo *cluster*
2. Calcular la separación del *cluster* $b^{(i)}$ al *cluster* más cercano como la distancia entre un punto en específico y todos los otros puntos del *cluster* más cercano
3. Calcular *silhouette* $s^{(i)}$ como la diferencia entre la cohesión y la separación del *cluster*, dividido por el máximo entre ambas mediciones (Ecuación 2.4).

Ecuación 2.4

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)},\ a^{(i)}\}}$$

Clustering c-means

c-means o *fuzzy c-means* es un algoritmo de *clustering* similar a *k-means*, pero se diferencian en que *c-means* asigna una instancia a uno o más grupos, con el objetivo de mejorar *k-means* y da flexibilidad al set de datos al pertenecer a distintos grupos. Trabaja con una asignación gradual de puntos medidos en grados en [0,1] (Raschka & Mirjalili, 2019; Valente De Oliveira & Pedrycz, 2007).

Tabla 5: Comparación de la asignación de instancias en k=3 *clusters* para *k-means* y *c-means*

k-means	***c-means***
$[x \in \mu^{(1)} \rightarrow w^{(i,j)} = 0]$	$[x \in \mu^{(1)} \rightarrow w^{(i,j)} = 0.1\ ]$
$[x \in \mu^{(2)} \rightarrow w^{(i,j)} = 1]$	$[x \in \mu^{(2)} \rightarrow w^{(i,j)} = 0.85]$
$[x \in \mu^{(3)} \rightarrow w^{(i,j)} = 0]$	$[x \in \mu^{(3)} \rightarrow w^{(i,j)} = 0.05]$

Fuente: (Raschka & Mirjalili, 2019)

Como se observa en la Tabla 5, el algoritmo *k-means* asigna una instancia a un grupo, otorgando puntuación 1 a la instancia perteneciente a un centroide y puntuación 0 a la instancia que no está asignada a un centroide. En *c-means* se otorgan puntuaciones entre 0 y 1 a cada instancia, siendo esta puntuación la probabilidad de pertenecer a un grupo en particular y la suma de las probabilidades de cada instancia es igual a 1 (Raschka & Mirjalili, 2019).

El algoritmo de *c-means* se describe como:

1. Se especifica el número *k* centroides los que serán los centros iniciales y se otorga aleatoriamente una puntuación a cada instancia
2. Se calculan los centroides para cada grupo (fórmula 2.5)
3. Se actualiza la puntuación de cada instancia para ser asignada a uno o más *clusters*
4. Se repiten los pasos 2 y 3 hasta que la puntuación de cada instancia no cambia o hasta que se alcance la cantidad máxima de iteraciones definidas por el usuario

El centroide se calcula (Ecuación 2.5):

$$\mu_j = \frac{\sum_{i=1}^{n} w^{(i,j)^m} x^{(i)}}{\sum_{i=1}^{n} w^{(i,j)^m}}$$ Ecuación 2.5

En donde:

μ_j: centroide del *cluster j*

$w^{(i,j)}$: puntuación asignada a la instancia perteneciente al *cluster j*

$x^{(i)}$: instancia perteneciente al *cluster j*

m : coeficiente *fuzziness*, por lo general es 2, controla el grado de difusión

La actualización de la puntuación se calcula con la Ecuación 2.6:

$$w^{(i,j)} = \left[\sum_{c=1}^{k} \left(\frac{\left|\left| x^{(i)} - u^{(j)} \right|\right|^2}{\left|\left| x^{(i)} - u^{(c)} \right|\right|^2} \right)^{\frac{2}{m-1}} \right]^{-1}$$ Ecuación 2.6

En donde:

$w^{(i,j)}$: puntuación asignada a la instancia perteneciente al *cluster j*

k : número de *clusters*

$x^{(i)}$: instancia perteneciente al *cluster j*

μ_j: centroide del *cluster j*

$\mu^{(c)}$: centroide de cada *cluster j*

La calidad del *cluster* se determina a través del coeficiente *silhouette*, al igual que en *k-means*, calculando en primer lugar la cohesión y la separación de cada *cluster*.

Gaussian Mixtures

Un modelo de mezcla gaussiana o *Gaussian mixture model*, es un modelo de probabilidad que asume que las instancias fueron generadas desde una mezcla de varias distribuciones gaussianas cuyos parámetros son desconocidos. Un *cluster* está formado por todas las instancias generadas de una sola distribución gaussiana, pero no es conocida la distribución o que parámetros se consideran, y cada instancia puede pertenecer a más de un *cluster* (Géron, 2019).

El algoritmo de *Gaussian mixture* se basa en el algoritmo *Expectation-Maximization*, algoritmo que se divide en dos partes, asignando en primer lugar las instancias en un *cluster*, parte *expectation*, y actualizando el *cluster*, parte *maximization*. Se tienen k clusters o funciones distribución gaussiana para el total de las instancias, con medias para cada *cluster* $\mu_1, \mu_2 \dots \mu_k$, y covarianzas para cada *cluster* $\Sigma_1, \Sigma_2 \dots \Sigma_k$, y la densidad de la distribución para cada *cluster* está representado por $\pi_1, \pi_2 \dots \pi_k$. Por lo tanto, el algoritmo para *Gaussian mixture* es (Ren & Mackay, 2019):

1. Iniciación de π, μ, Σ
2. Evaluación de la función de probabilidad logarítmica
3. Iniciación del ciclo, parte *expectation*
4. Re-estimación de los parámetros π, μ, Σ, parte *maximization* (Ecuación 2.8, 2.9 y 2.10)
5. Re-evaluación de la función de probabilidad logarítmica (Ecuación 2.11)
6. Verificar convergencia, si cumple con error establecido se termina el ciclo, de lo contrario, repetir pasos 3, 4 y 5

Probabilidad de que pertenezca a cada distribución o *cluster* está dada por la Ecuación 2.7*:*

$$\gamma_k(X) = \frac{\pi_k \mathcal{N}(x_i;\, \mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j \mathcal{N}(x_i;\, \mu_j, \Sigma_j)} \qquad \text{Ecuación 2.7}$$

En donde:

$\gamma_k(X)$: probabilidad de que una instancia pertenezca al *cluster* k

π_k: peso de la instancia en el *cluster* k

π_j: peso de la instancia en cada *cluster* k

N_k: número de instancias asignadas al *cluster* k

N: número total de instancias

μ_k : media de la distribución de la instancia en el *cluster* k

μ_j: peso de la instancia en cada *cluster* k

Σ_k: covarianza del *cluster* k

Σ_j: covarianza de cada *cluster* k

Media del *cluster*:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(x_n) \cdot x_n \qquad \text{Ecuación 2.8}$$

En donde:

$\gamma_k\,(x_n)$: probabilidad de cada instancia de pertenecer al *cluster* k

x_n: instancia n

Covarianza del *cluster:*

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_k\,(x_n)(\,x_n - \mu_k)^T(\,x_n - \mu_k) \qquad \text{Ecuación 2.9}$$

Densidad o peso del *cluster*:

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{n=1}^{N}\gamma_k\,(x_n)}{N} \qquad \text{Ecuación}$$

Función de probabilidad logarítmica:

$$ln\ p(X;\,\pi,\mu,\Sigma) = \sum_{n=1}^{N}\ln\left(\sum_{k=1}^{K}\pi_k\,\mathcal{N}(x_i;\,\mu_k,\,\Sigma_k)\right) \qquad \text{Ecuación}$$

En resumen, el método *Gaussian mixture* a través del algoritmo *expectation-maximization* inicializa los parámetros del *cluster* aleatoriamente, asigna instancias al *cluster* (expectación) y calcula la probabilidad de que cada instancia pertenezca al cluster, luego actualiza el *cluster* (maximización) utilizando todas las instancias del set de datos y calcula que cada instancia verdaderamente pertenezca al *cluster* asignado, por lo que no solo calcula el centro de cada cluster ($\mu_1$hasta μ_k), sino que también calcula el tamaño y orientación ($\Sigma_1$hasta Σ_k) y los pesos relativos ($\pi_1$hasta π_k) (Géron, 2019).

Eligiendo el número de *clusters*

En *gaussian mixture* no es posible utilizar la inercia o *silhoutte* para calcular el número de grupos ya que no son confiables cuando los *clusters* no son esféricos o con diferentes tamaños, por lo que, para esta tarea, se utiliza el criterio de información Bayesiana (BIC por su sigla en inglés *Bayesian information criterion*) o el criterio de información Akaike (AIC, *Akaike information criterion*) (Ecuación 2.12):

$$BIC = \log(m)\, p - 2\log(\hat{L})$$
$$AIC = 2p - 2\log(\hat{L})$$

Ecuación

En donde:

m : número de instancias

p : número de parámetros aprendidos por el modelo

$\hat{L}$: valor maximizado de la función distribución del modelo

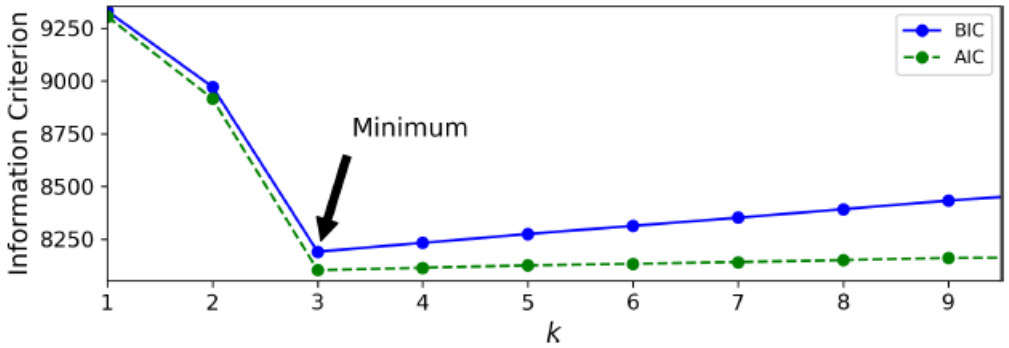


Figura 8: comparación de BIC y AIC con distintos números de *clusters*. Fuente: (Géron, 2019)

BIC y AIC deben tomar el valor más bajo posible, como se observa en la Figura 8, así se asegura el número correcto de *clusters* (Géron, 2019).

2.2.5. Aprendizaje Supervisado

Los algoritmos de aprendizaje operan de la misma manera que con los humanos, con la guía proporcionada de cómo es el mundo y cómo difiere de la suposición actual del algoritmo. Para lograr esto, es esencial etiquetar previamente los datos (Vaughan, 2020). Un ejemplo común de aprendizaje supervisado es la clasificación, como en el caso de un filtro de spam. Para entrenarlo, se utilizan numerosos ejemplos de correos electrónicos previamente etiquetados como spam o no spam, y el algoritmo aprende a clasificar correctamente nuevos correos electrónicos en base a esta información (Gerón, 2019). Gracias a esta etiquetación, el algoritmo tiene la capacidad de comparar su predicción más reciente con la realidad y ajustarse en consecuencia (Vaughan, 2020).

Ensembled Learning

Ensemble Learning, que es una técnica en la que se combinan las predicciones de varios modelos predictivos para obtener mejores resultados que los que podría lograr un único predictor (Géron, 2019).

Un ejemplo de esto es el *Random Forest*, que utiliza múltiples árboles de decisión entrenados en diferentes subconjuntos de datos y luego combina sus predicciones para tomar una decisión final. A pesar de su simplicidad, *Random Forest* es uno de los algoritmos de Aprendizaje Automático más poderosos (Géron, 2019).

Ensembled Learning emplea la técnica de clasificadores de votación, que involucra entrenar varios clasificadores individuales, como Regresión Logística, SVM, *Random Forest*, *K-Nearest Neighbors*, entre otros (como se muestra en la Figura 9), y luego combinar sus predicciones para determinar la clase final. Esta técnica se llama "clasificador de votación duro" y la clase que obtiene la mayoría de los votos se selecciona como la predicción final, como se ilustra en Figura 10 (Géron, 2019).

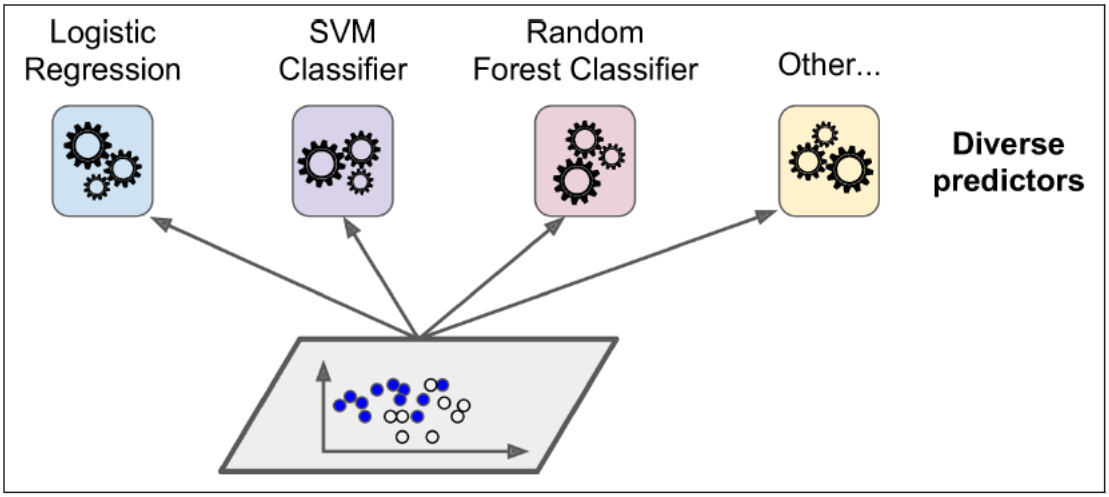


Figura 9: Esquema de entrenamiento de distintos clasificadores. Fuente: (Géron, 2019)

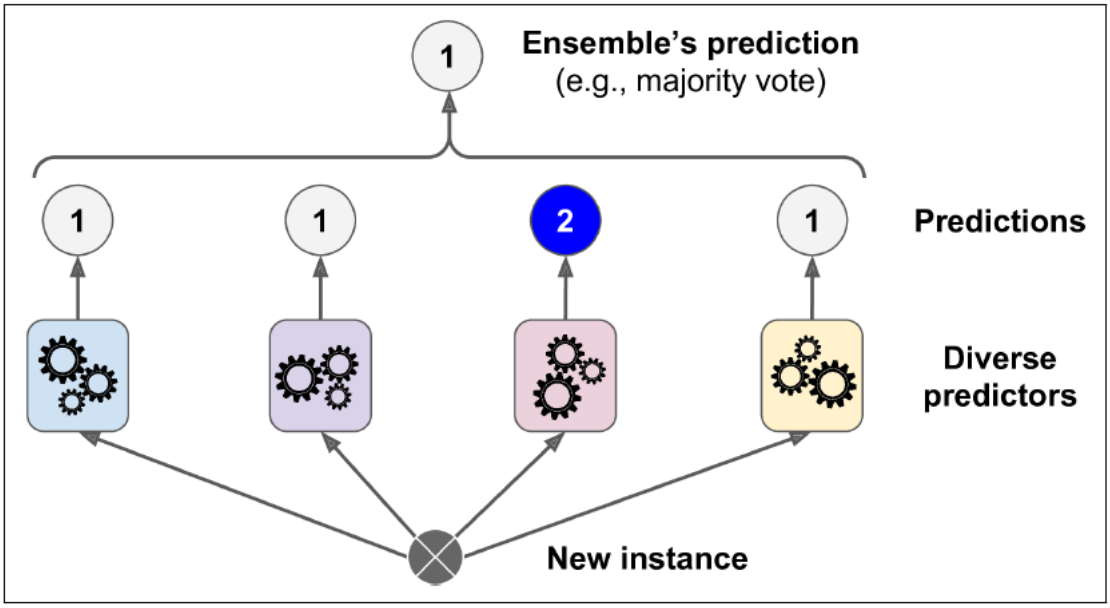


Figura 10: esquema de técnica "clasificador de votación duro". Fuente: (Géron, 2019)

Lo que hace que esta técnica sea interesante es que a menudo supera en precisión al mejor clasificador individual en el conjunto. Incluso si cada clasificador individual es relativamente débil y solo tiene un rendimiento ligeramente mejor que el azar, el conjunto de clasificadores puede convertirse en un clasificador fuerte cuando se combinan adecuadamente (Géron, 2019).

Una analogía útil para comprender esto es imaginar una moneda ligeramente sesgada que tiene un 51% de probabilidad de caer cara y un 49% de probabilidad de caer cruz. Si lanzamos esta moneda 1,000 veces, es probable que obtengamos alrededor de 510 caras y 490 cruces, lo que significa una mayoría de caras. A medida que aumenta el número de lanzamientos, la probabilidad de obtener una mayoría de caras se acerca cada vez más al 75%. Esto se debe a la ley de los grandes números, que establece que a medida que repetimos un experimento muchas veces, la proporción de resultados se acerca a las probabilidades subyacentes (Géron, 2019).

En resumen, los clasificadores de votación se benefician de la diversidad de enfoques de varios clasificadores débiles, lo que les permite obtener un rendimiento sólido y superar las limitaciones de un solo clasificador (Géron, 2019).

Bagging

El *bagging* (*Bootstrap Aggregating*) es un método de aprendizaje en conjunto que se basa en el muestreo de arranque. Su funcionamiento se basa en entrenar varios predictores utilizando conjuntos de datos diferentes, generados mediante el muestreo aleatorio con reemplazo o sin reemplazo del conjunto de entrenamiento original. Cada predictor se entrena de forma independiente en uno de estos subconjuntos de datos (Gerón, 2019).

Una característica clave del *bagging* es que, debido al muestreo con reemplazo, las instancias de entrenamiento pueden aparecer varias veces en cada subconjunto. Esto significa que cada predictor puede aprender de las mismas instancias de entrenamiento más de una vez, lo que aumenta la diversidad entre los predictores (Gerón, 2019).

Una vez que todos los predictores están entrenados, el conjunto realiza predicciones para una nueva instancia agregando las predicciones individuales de cada predictor. En el caso de clasificación, se utiliza la moda estadística (la predicción más frecuente) como función de agregación, mientras que en la regresión se utiliza el promedio de las predicciones (Gerón, 2019).

A pesar de que cada predictor individual tiene un sesgo más alto que si se entrenara en el conjunto de entrenamiento original, la agregación de las predicciones de múltiples predictores reduce tanto el sesgo como la varianza. Esto conduce a un conjunto que generalmente tiene un sesgo similar pero una varianza menor en comparación con un solo predictor entrenado en el conjunto de entrenamiento original (Gerón, 2019).

El *bagging* se utiliza para mejorar la precisión y la robustez de los modelos de aprendizaje automático y es especialmente eficaz cuando se combinan múltiples predictores que son diversificados en términos de sus entrenamientos. Además, el proceso de entrenamiento y predicción se puede llevar a cabo en paralelo, lo que hace que el *bagging* sea altamente escalable y adecuado para problemas que requieren un alto poder de cómputo (Gerón, 2019).

El algoritmo de *Bagging* se basa en un método de aprendizaje en conjunto que se basa en el muestreo de arranque. Funciona tomando un conjunto de datos con "m" muestras y generando repetidamente un nuevo conjunto de datos tomando una muestra al azar del original y copiándola al nuevo conjunto. Este proceso se repite "m" veces para crear un conjunto de datos de tamaño "m", donde algunas muestras originales pueden aparecer varias veces y otras no aparecer en absoluto. Este proceso se repite "*T*" veces para crear

"*T*" conjuntos de datos, cada uno con "m" muestras. Luego, se entrenan modelos base en estos conjuntos de datos y se combinan. La combinación de las predicciones de los modelos base se realiza mediante votación simple en tareas de clasificación o promedio simple en tareas de regresión (Ecuación 2.13), y todos los modelos base tienen el mismo peso en esta combinación (Zhou, 2021).

En donde:
Conjunto de entrenamiento: *D* = {(*x1, y1*), (*x2, y2*), …, (*xm, ym*)};
Algoritmo de aprendizaje en Base $\mathcal{L}$;
Número de rondas de entrenamiento *T*.

Proceso:

1. **for** $t = 1$ to T **do**
2. $h_t = \mathcal{L}(D, D_{bs})$
3. **end for**

$$H(x) = \arg\max_{y \epsilon Y} \sum_{t=1}^{T} \mathrm{II}(h_t(x) = y) \qquad \text{Ecuación 2.13}$$

Random Forest

Random Forest es un conjunto de Árboles de Decisión, generalmente entrenados mediante el método de *bagging*, que en lugar de buscar la mejor característica absoluta al dividir un nodo, como lo hacen los árboles de decisión convencionales, *Random Forestm* busca la mejor característica entre un subconjunto aleatorio de características. Esta aleatoriedad aumenta la diversidad entre los árboles en el conjunto, lo que, en términos generales, disminuye la varianza del modelo. Aunque cada árbol individual puede tener un cierto grado de sesgo, la combinación de múltiples árboles generalmente produce un modelo global más preciso y robusto (Géron, 2019).

Redes Neuronales

La investigación sobre redes neuronales se remonta a un período largo y se ha convertido en un campo interdisciplinario amplio. Aunque las redes neuronales tienen diversas definiciones en diferentes disciplinas, este libro adopta una definición ampliamente aceptada: las redes neuronales artificiales son redes masivamente paralelas de elementos

simples, diseñadas para interactuar con objetos del mundo real de manera similar a los sistemas nerviosos biológicos (Zhou, 2021).

El elemento fundamental en las redes neuronales es la neurona o unidad, que, cuando se excita, envía señales a otras neuronas interconectadas para cambiar sus potenciales eléctricos. Estas señales son activadas cuando el potencial eléctrico supera un umbral. Este proceso se basa en el modelo McCulloch-Pitts, que se utiliza para describir el funcionamiento de las neuronas y todavía se emplea en la actualidad. Como se ilustra en la Figura 11, cada neurona recibe señales de entrada desde otras n neuronas a través de conexiones ponderadas. Estas señales de entrada se suman, y el resultado se compara con un umbral. Si la suma supera este umbral, la neurona se activa y envía una señal de salida. Esta activación se logra mediante una función de activación, que puede ser una función de paso o una función sigmoide, como se ilustra en la Figura 12.a) La función de paso mapea la entrada a valores binarios (0 o 1), mientras que la función sigmoide ajusta la entrada en un intervalo abierto entre 0 y 1, como se ilustra Figura 12.b) (Zhou, 2021).

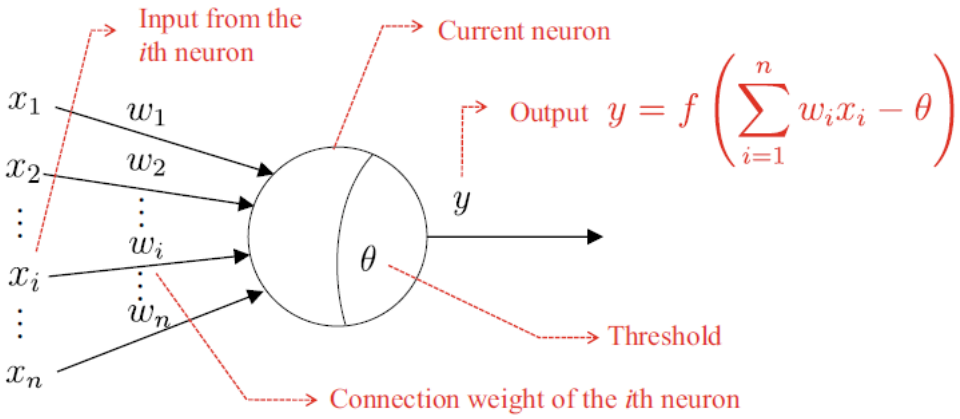


Figura 11: Modelo de neurona M-P. Fuente: (Zhou, 2021).

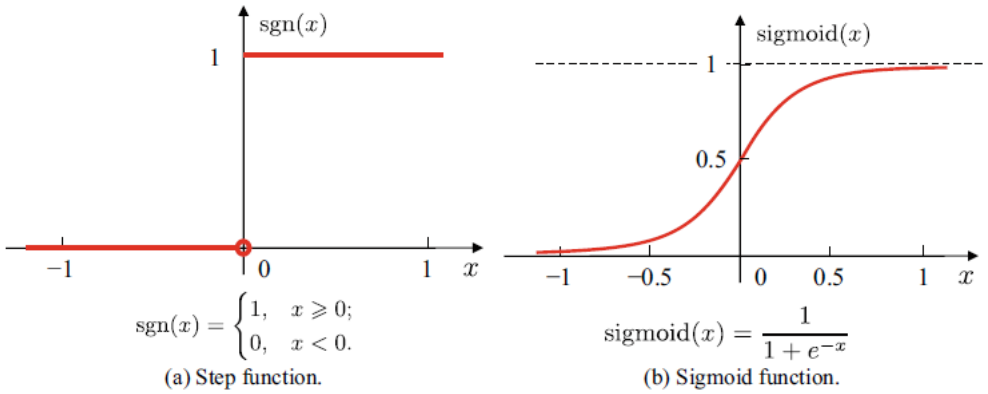


Figura 12: Funciones de activación típicas de neuronas. a) función paso. b) función sigmoide. Fuente: (Zhou, 2021)

Una red neuronal se forma al conectar estas neuronas en una estructura de capas. Desde una perspectiva informática, una red neuronal es vista como un modelo matemático con muchos parámetros, y su similitud con las redes biológicas a veces se considera secundaria, como por ejemplo la función de anidado múltiple $y_j = f(\sum_i w_i x_i - \theta_j)$. Los algoritmos de aprendizaje de redes neuronales son respaldados por pruebas matemáticas (Zhou, 2021).

Perceptrón

El perceptrón es un clasificador binario que consta de dos capas de neuronas, como se muestra en la Figura *13*. La capa de entrada recibe señales externas y las transmite a la capa de salida, que es una neurona M-P, también conocida como unidad lógica de umbral. El perceptrón puede realizar fácilmente las operaciones lógicas "Y", "O" y "NO". Si asumimos que la función f en $y_j = f(\sum_i w_i x_i - \theta)$. es la función de paso que se muestra en la figura 12, las operaciones lógicas se pueden llevar a cabo de la siguiente manera:

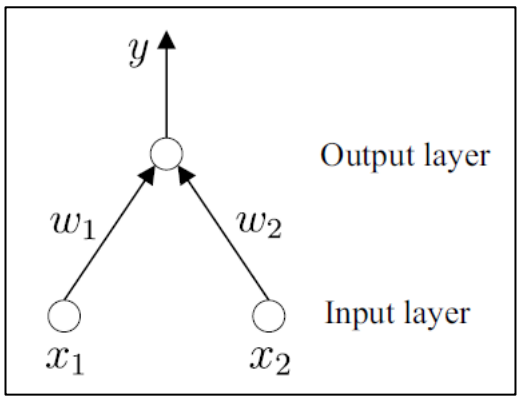


Figura 13: Un perceptrón con dos neuronas de entrada. Fuente: (Zhou, 2021).

- "***AND***" ($x_1 \wedge x_2$): si $w_1 = w_2 = 1, \theta = 2$, entonces $y = f(1 . x_1 + 1 . x_2 - 2)$, *and* $y = 1$ si *and* si y solo si $x_1 = x_2 = 1$;
- "***OR "*** ($x_1 \vee x_2$): si $w_1 = w_2 = 1, \theta = 0.5$, entonces $y = f(1 . x_1 + 1 . x_2 - 0.5)$, *and* $y = 1$ cuando $x_1 = 1$ *or* $x_2 = 1$;
- ***"NOT"*** ($\neg x_1$): si $w_1 = -0.6, w_2 = 0, \theta = -0.5$, entonces $y = f(-0.6 . x_1 + 0 . x_2 - 0.5)$, *and* $y = 0$ cuando $x_1 = 1$ *and* $y = 1$ cuando $x_1 = 0$.(Zhou, 2021)

De manera más general, los pesos $w_i (i = 1, 2, \ldots, h)$ y el umbral θ pueden ser ajustados utilizando datos de entrenamiento. Si consideramos el umbral θ como un componente adicional con el peso de conexión w_{n+1} y una entrada fija de -1.0, entonces los ajustes de peso y umbral se simplifican en un proceso unificado. El aprendizaje del perceptrón es sencillo: para una muestra de entrenamiento (x, y), si el perceptrón produce una salida $\hat{y}$, entonces el peso es actualizado por la Ecuación 2.14:

$$\Delta w_i \leftarrow w_i + \Delta w_i,$$

Ecuación 2.14

$$\Delta w_i = \eta (y - \hat{y}) x_i$$

En donde:
x_i: corresponde al valor del i-ésimo neurona de entrada
η: generalmente se establece como un número positivo pequeño, por ejemplo, 0.1 (Zhou, 2021).

La tasa de aprendizaje η, que está en el rango (0, 1), se conoce como la velocidad de aprendizaje. A partir de la ecuación $\Delta w_i \leftarrow w_i + \Delta w_i$, podemos observar que el perceptrón permanece sin cambios si predice correctamente la muestra (x, y) (es decir, $\hat{y} = y$). De lo contrario, el peso se actualiza en función del grado de error (Zhou, 2021).

La capacidad de aprendizaje de los perceptrones es bastante limitada debido a que solo la capa de salida tiene funciones de activación, es decir, solo una capa de neuronas funcionales. De hecho, los problemas de "AND", "OR" y "NOT" son todos problemas linealmente separables. Minsky y Papert (1969) demostraron que debe existir un hiperplano lineal que pueda separar dos clases si son linealmente separables. Esto significa que el proceso de aprendizaje del perceptrón está garantizado para converger hacia un vector de pesos apropiado $w = (w_1; w_2; \ldots; w_{n+1})$, como se ilustra en la figura 14. De lo contrario, se producirán fluctuaciones en el proceso de aprendizaje y no se podrá encontrar una solución

adecuada, ya que w no puede estabilizarse. Por ejemplo, el perceptrón no puede resolver ni siquiera problemas simples no linealmente separables como "XOR", como se muestra en la Figura *14*. (Zhou, 2021).

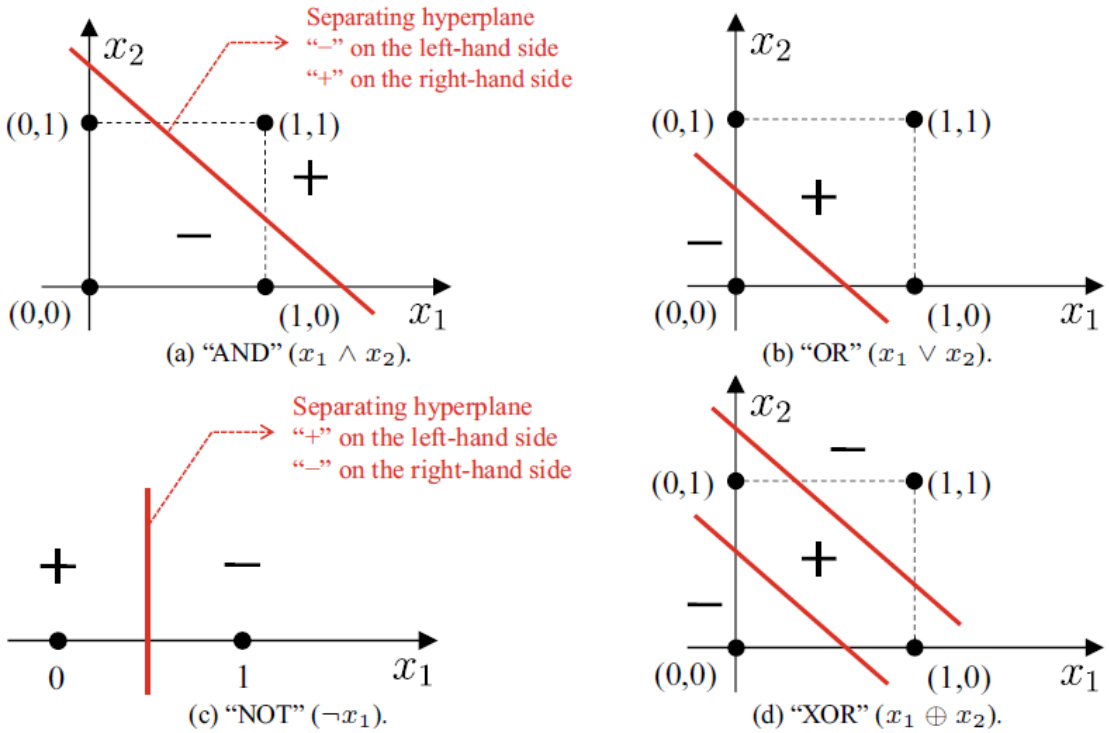


Figura 14: 'AND', 'OR', y 'NOT' son problemas que pueden separarse mediante una línea recta, mientras que 'XOR' es un problema que no se puede separar de manera lineal. Fuente: (Zhou, 2021).

Para resolver problemas que no se pueden separar de manera lineal, podemos utilizar neuronas funcionales de múltiples capas. Por ejemplo, el sencillo perceptrón de dos capas ilustrado en la Figura 15 puede resolver el problema del 'XOR'. En la Figura 15.a), la capa de neuronas entre la capa de entrada y la capa de salida se conoce como la capa oculta y tiene funciones de activación similares a las de la capa de salida (Zhou, 2021).

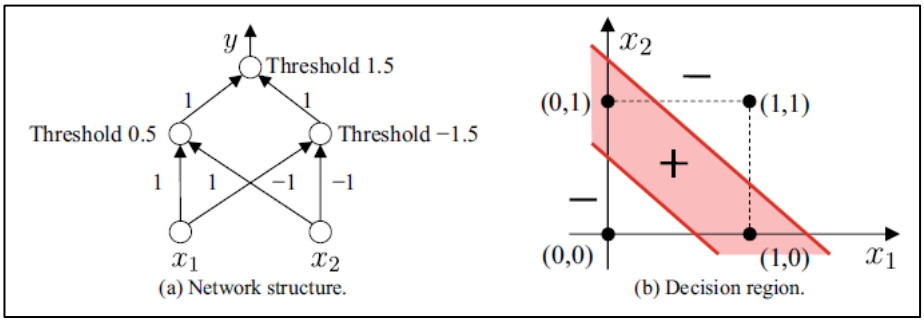


Figura 15: Un perceptrón de dos capas que resuelve el problema del 'XOR'. Fuente: (Zhou, 2021).

La Figura 16 muestra dos ejemplos de estructuras de redes neuronales multicapa en las que cada capa se conecta completamente con la siguiente. Sin embargo, las neuronas dentro de la misma capa o en capas no adyacentes no están conectadas. Estas redes neuronales siguen una estructura conocida como "redes neuronales de avance multicapa". En este tipo de redes, la capa de entrada recibe señales externas, las capas ocultas y de salida procesan estas señales, y la capa de salida emite las señales procesadas. Básicamente, la capa de entrada solo recibe datos de entrada sin realizar ningún procesamiento, mientras que las capas ocultas y de salida contienen neuronas funcionales. Cuando solo hay dos capas funcionales, la red neuronal de la figura 16.a) a menudo se denomina "red neuronal de dos capas", aunque también se denomina "red neuronal con una sola capa oculta" para evitar confusiones. Cuando una red neuronal tiene al menos una capa oculta, la llamamos "red neuronal multicapa". El proceso de aprendizaje en las redes neuronales implica ajustar los pesos de conexión entre las neuronas y los umbrales de las neuronas funcionales a partir de los datos de entrenamiento. En esencia, el "conocimiento" adquirido por las redes neuronales se refleja en estos pesos de conexión y umbrales.

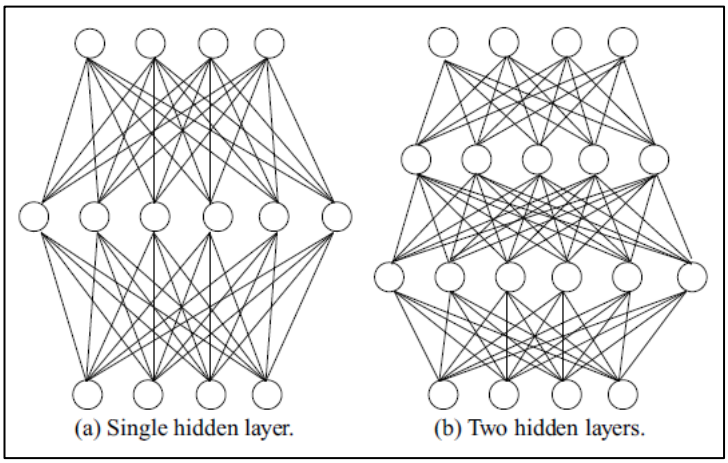


Figura 16: Estructura de red neuronal multicapa. a) red neural con una capa oculta. b) red neuronal con dos capas ocultas. Fuente: (Zhou, 2021).

2.2.6. Aprendizaje Semi Supervisado basado en grafos

Un grafo es un conjunto de objetos llamados nodos o vértices y están unidos entre ellos con aristas llamadas relaciones o conexiones. Estos nodos representan entidades y las aristas corresponde a la relación que tienen los nodos entre sí, las cuales pueden ser directas o indirectas, pero siempre tienen un punto de partida y llegada, y también pueden contener

propiedades, tal como se muestra en la Figura 17. Los nodos pueden tener propiedades (pares clave-valor), pueden tener una o más etiquetas y cuentan con un número diferente de nodos vecinos (Needham & Hodler, 2019; Robinson et al., 2015).

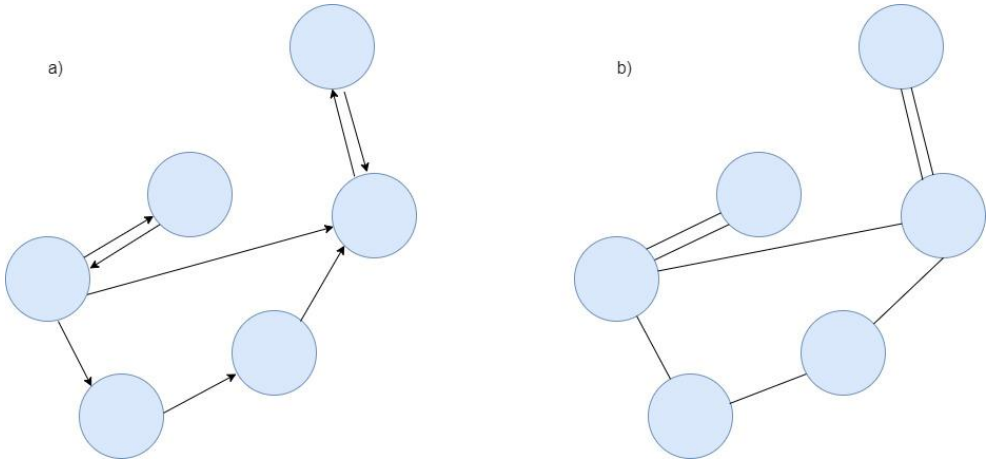


Figura 17: a) Grafo con 6 nodos y 8 conexiones directas, b) grafo con 6 nodos y 8 conexiones indirectas Fuente: Elaboración propia.

Técnicas como *clustering*, se basan en la distancia Euclidiana, definiendo un centroide y los datos más cercanos a un centroide se agrupan, por lo que los datos son representados en el espacio Euclidiano, sin embargo, existen estructuras de datos que no se rigen por esta distancia, sino que se generan desde espacios no-Euclidianos y no siguen algún tipo de orden, como datos generados por redes sociales, expresión de genes, neurociencia, fraudes, entre otros (Asif et al., 2021; Bronstein et al., 2016; Z. Wu et al., 2019).

Estas estructuras son representadas a través de grafos con complejas relaciones e interdependencia entre los objetos, y ya que no pueden ser modeladas a través de métodos de *deep learning* como *clustering*, clasificación, regresión, u otras, es que se han desarrollado métodos como redes neuronales de grafos (GNN por sus siglas en ingles *Graph Neural Networks*) (Sanchez-Lengeling et al., 2021; Wu et al., 2019).

Para un set de datos, cada instancia corresponde a un nodo y estos se conectan de acuerdo con la correlación de estas instancias, la fuerza de una conexión indica el grado de similaridad y los nodos pueden estar o no etiquetados. Por lo tanto, un grafo indirecto está dado por la Ecuación 2.15:

$$G = (V, E) \qquad \text{Ecuación 2.15}$$

En donde:

V: set de N nodos $v_i \in V$

v_i: notación del nodo i

E: set de líneas $(v_i, v_j)\ \epsilon\ \mathcal{E}$

(v_i, v_j): notación de la unión de dos nodos

El vecindario de un nodo está dado por (Ecuación 2.16):

$$N(v) = \{u\ \in V | (v, u)\ \in, E\}$$ Ecuación 2.16

Además, cuenta con una matriz adyacente binaria o con pesos definida A de dimensión $n\ x\ n$ con $n\ \epsilon\ \mathbb{R}$ nodos:

$$A\ \in\ \mathbb{R}^{nxn}$$

En donde:

n: número de nodos

Matriz de atributos del grafo:

$$X\ \epsilon\ \mathbb{R}^{dxd}$$

En donde:

d: número de atributos del nodo

Clasificación de nodos

La clasificación de nodos de un grafo es un tipo de aprendizaje no supervisado y tiene como fin predecir etiquetas basado en los atributos y estructura del vecindario de los nodos. Para llevar a cabo esta tarea, se utilizan redes neuronales de grafos. El modelo GNN se considera una función $f(X, A)$, condicionada por la matriz de atributos A y la matriz adyacente A (Kipf & Welling, 2017; Maurya et al., 2023).

Redes neuronales de grafos

Las redes neuronales de grafos o GNN (*Graph Neural Network*) siguen las estrategias de *deep learning* y se utilizan para modelos como clasificación y predicción con sets de datos que no siguen la estructura Euclidiana. GNN son un grupo de modelos de redes neuronales que están diseñadas para varias tareas. Aprovechan el mecanismo de propagación de atributos para agregar información a los nodos vecinos y utilizar transformaciones no lineales con una matriz de pesos entrenable para obtener los encajes finales de los nodos (Figura 18) (Asif et al., 2021; Z. Wu et al., 2019).

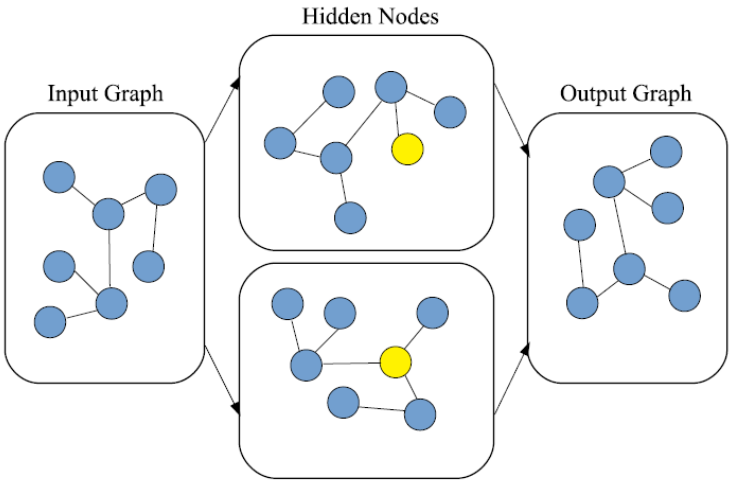


Figura 18: Red neuronal de grafos. Capa de entrada con el grado inicial, en las capas internas ocurre la tarea a realizar (clasificación, predicción, entre otras), capa de salida con resultado. Fuente: (Asif et al., 2021)

GNN se propaga como (Ecuación 2.17):

$$X_v^{(\ell+1)} = f_0^{\ell+1}(X_v^{\ell}, \{X_u^{\ell} : u \epsilon N(v)\}) \qquad \text{Ecuación 2.17}$$

En donde:

$X_v^{(\ell+1)}$: matriz de atributos

X_v^{ℓ}: matriz de atributos anterior

X_u^{ℓ}: matriz de atributos del vecindario

Redes neuronales de grafos convolucionales

Las redes neuronales convolucionales o CNN (por sus siglas en inglés *Convolutional Neural Network*) son arquitecturas del *deep learning* y se utilizan para identificar patrones para detectar objetos, clases o categorías. Se utilizan principalmente para reconocimiento de imágenes, reconocimiento de voz y señales, y en los últimos años se han utilizado para predicción o clasificación de grafos (nodos o grafos completos). Están compuestas con una capa de entrada, una de salida, y capas intermedias convolucionales, todas activadas por funciones de activación (Géron, 2019).

La convolución es una operación matemática en donde dos funciones producen una tercera función describe la superposición entre ambas y se define como la integral del producto de ambas funciones (Géron, 2019).

La red neuronal de grafos convolucional o GCN (por sus siglas en inglés *Graph Convolutional Neural Network*), genera una representación del nodo v agregando sus propios atributos x_v y los atributos de los nodos vecinos x_u, donde $u \epsilon N(v)$. GCN apila múltiples capas convolucionales para extraer una representación de mayor nivel. Cada capa encapsula la representación oculta del cada nodo agregando información de los atributos de los nodos vecinos, luego, una transformación no linear (función de activación) es aplicada para generar una salida (resultado), como se muestra en la Figura 19 (Z. Wu et al., 2019).

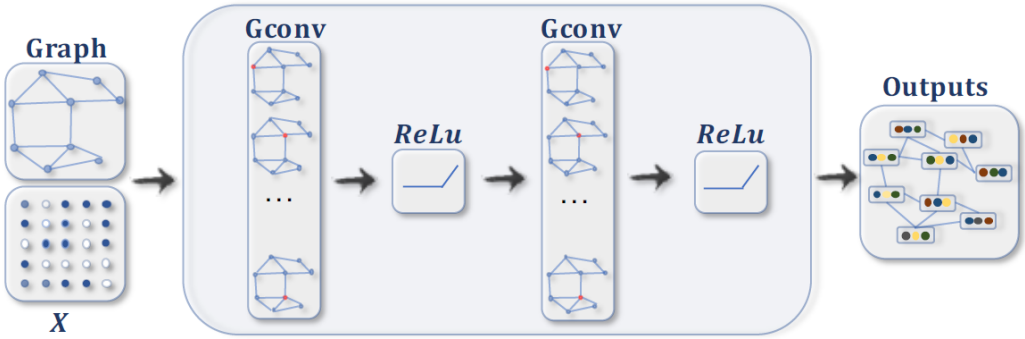


Figura 19: Estructura de una red neuronal de grafos convolucional. Fuente: (Wu et al., 2019)

Una capa de GCN se propaga como (Ecuación 2.18):

$$X_v^{\ell+1} = W^{\ell+1} \sum_{w \epsilon N(v) \cup \{v\}} \frac{1}{c_{w,v}} \cdot X_v^{\ell} \qquad \text{Ecuación 2.18}$$

En donde:

$X_v^{\ell+1}$: matriz de atributos

$W^{\ell+1}$: matriz de pesos aprendida

$c_{w,v}$: coeficiente de normalización

X_v^{ℓ}: matriz de atributos anterior

3. Objetivos concretos y metodología de trabajo

A continuación, se presentan el objetivo general y los objetivos específicos a realizar en el presente trabajo de fin de máster.

Para esto, se hará uso de las mediciones de parámetros físicos, químicos y biológicos obtenidos en muestreos *in situ* en los diferentes sistemas hidrográficos de interés definidos en la Resolución Exenta 1854/2022 de la Subsecretaría de Pesca y Acuicultura del Ministerio de Economía, Fomento y Turismo.

3.1. Objetivo general

Como objetivo general del presente trabajo es examinar y comparar diferentes modelos de *custering* y posteriormente, aplicar modelos de aprendizaje supervisado para predecir factores como la calidad de agua y condiciones que favorecen la proliferación y supervivencia de Didymo en Chile. Estos modelos serán aplicados en la detección futura de Didymo por el Instituto de Fomento Pesquero de Puerto Montt, Chile.

3.2. Objetivos específicos

- Limpiar, analizar, interpretar y normalizar a través de un análisis exploratorio de datos (EDA) los datos registrados en estudios sobre los factores ambientales químicos, físicos y biológicos involucrados en la proliferación de Didymo en Chile para obtener un set de datos uniforme para su posterior clasificación.
- Clasificar datos registrados *in situ* de las condiciones que favorecen la proliferación y supervivencia de Didymo en los diferentes tipos de sistemas hidrográficos de Chile a través de modelos de *clustering* como k-means, c-means y bayesiano.
- Aplicar y comparar modelos de aprendizaje supervisado como *Ensembled Learning* y Redes Neuronales en base a los resultados obtenidos del modelo de *clustering* para predecir la presencia de Didymo en los sistemas hidrográficos de Chile.
- Entrenar y evaluar los modelos implementados para comparar métricas de rendimiento y definir el modelo a usar en futuros datos recopilados por el Instituto de Fomento Pesquero de Puerto Montt, Chile.
- Generar un informe con los hallazgos conseguidos al aplicar técnicas de inteligencia artificial en la detección temprana de Didymo en Chile.

3.3. Metodología del trabajo

La metodología de trabajo se dividió en 3 partes: análisis exploratorio de datos, modelado y entrenamiento y análisis de resultados.

Para obtener los datos se debe realizar una solicitud a la Subsecretaría de Pesca y Agricultura del Gobierno de Chile a través de la Ley de Transparencia N°20.285.

3.3.1. Análisis exploratorio de datos (EDA)

3.3.1.1. Preprocesamiento de datos:

Se realizó el procesamiento de los datos en primer lugar:

1. Reemplazo de caracteres especiales

- En Microsoft Excel se realizó el reemplazó de los siguientes caracteres:
- Ñ → N
- á, é, í, ó, ú → a, e, i, o, u

2. Homogenización de nombres:

Se observó que el atributo “rio”, presenta categorías no homogéneas, es decir, el mismo río está escrito de distintas formas, por ejemplo: Río Truful Truful, Truful-Truful y Truful Truful, los 3 corresponden al mismo sitio. Para corregir este problema se utilizó la herramienta OpenRefine. A continuación, se explica cada paso del procedimiento:

2.1. Carga de datos

- Se cargó el archivo .xlsx en herramienta OpenRefine.

2.2. Agrupación

- Se desplegó el menú de la variable “rio”, se seleccionó “*Facet*” → “*Text Facet*”.
- En el menú lateral izquierdo que se abrió, se seleccionó “*Cluster*” y se verificó el nombre de los grupos detectados automáticamente por la herramienta. Se probaron distintos métodos y funciones para lograr el mayor agrupamiento posible.

- Los valores que no fueron posibles de agrupar automáticamente se agruparon manualmente.

2.3. Descarga de set de datos

- El set de datos corregido se descargó en formato .csv.

3.3.1.2. EDA:

El procedimiento aplicado en el código de "*Exploratory Data Analysis* (EDA)" tiene como objetivo principal de este análisis fue preparar y limpiar un conjunto de datos para su posterior análisis. A continuación, se explica cada paso del procedimiento:

1. Carga de Datos:

- Se utilizó la biblioteca Pandas para cargar los archivos llamados "Ambientales.csv" y “Diatomeas.xlsx”. Se especificó la codificación como “ISO-8859-1” y el delimitador como “;”.

2. Copia de Columna:

- Se copió la columna “*Didymosphenia geminata*” del dataset “Diatomeas” al dataset “Ambientales”.

3. Reemplazo de Valores:

- Se identificaron ciertos valores, como: “s/m”; “s/i”; “s/n”; “s/h”; “ds/m”; yy “-”, que no eran válidos en el conjunto de datos y se reemplazaron por NaN (valores nulos) utilizando la biblioteca NumPy.

4. Limpieza de Caracteres Especiales:

- Se recorrieron todas las columnas y se eliminó el carácter “<“, al principio de los valores en cada campo.

5. Transformación de Tipo de Datos:

- Se transformaron las columnas de tipo “*object*”, en columnas de tipo “*category*” para ahorrar memoria y acelerar el análisis.

6. Eliminación de Campos no Relevantes:

- Se eliminaron múltiples columnas que se consideraron no relevantes para el análisis, como información geoespacial, fechas, identificadores y otras variables.

7. Reemplazo de Comas por Puntos:

- Se reemplazaron las comas por puntos en todos los valores del conjunto de datos. Esto fue necesario para que los valores fueran interpretables como números decimales en lugar de cadenas de texto.

8. Cálculo de Medias para Datos por Río:

- Para un conjunto específico de columnas, se calculó el valor promedio por río y se rellenaron los valores nulos en esas columnas con el promedio correspondiente. Se creó un nuevo *DataFrame* temporal, se calculó la media y se fusionó con el *DataFrame* original.

9. Cálculo de Medias para Datos por Subcuenca:

- Para un conjunto específico de columnas, se calculó el valor promedio por subcuenca y se rellenaron los valores nulos en esas columnas con el promedio correspondiente. Se creó un nuevo *DataFrame* temporal, se calculó la media y se fusionó con el *DataFrame* original.

10. Cálculo de Medias para Datos por Cuenca:

- Para un conjunto específico de columnas, se calculó el valor promedio por cuenca y se rellenaron los valores nulos en esas columnas con el promedio correspondiente. Se creó un nuevo *DataFrame* temporal, se calculó la media y se fusionó con el *DataFrame* original.

11. Cálculo de Medias Globales para Datos:

- Se calcularon las medias globales para un conjunto de columnas y se rellenaron los valores nulos con las medias correspondientes.

12. Truncado de Decimales:

- Se realizaron ajustes en la cantidad de decimales de las columnas numéricas para homogeneizar la información según un estándar predefinido.

13. Visualización de Histogramas:

- Se generaron histogramas para visualizar la distribución de algunas de las variables, como “T°”, “P T”, “PO4”, “Si T”, “Ac_entorno_Pesca_deportiva” y “Clar_agua_Ligeramente Turbia”.

14. Visualización de Diagramas de Dispersión:

- Se generaron diagramas de dispersión para visualizar la distribución de las variables.

15. Visualización de Mapas de Calor:

- Se generaron mapas de calor, que muestran la correlación de las variables.

16. Creación de una columna

- Se creó la columna “categoría”, la cual combina las columnas: '%Cob_algal_ausente', '%Cob_algal_Pequenas colonias', %Cob_algal_Mediana', '%Cob_algal_Alta', '%Cob_algal_Muy Alta'.

Este procedimiento tenía como objetivo preparar el conjunto de datos para el análisis posterior, lo que incluyó la corrección de valores erróneos, la imputación de datos faltantes y la transformación de datos para su correcta interpretación. Además, se realizó una visualización inicial para comprender mejor la distribución de algunas variables clave. Este enfoque fue fundamental en la fase de exploración de datos antes de aplicar técnicas más avanzadas de análisis y modelado.

17. Eliminación de variables utilizadas para calcular datos faltantes

- Se eliminaron las variables 'cod_cuenca', 'cod_subcuenca', 'rio', 'epoca_muest', ya que el objetivo de estas era solo utilizarlas como referencia para completar los datos faltantes del set de datos.

18. Set de datos para etapas posteriores:

- Se guardó como archivo .csv el set de datos a utilizar en etapas posteriores, el que cuenta con las variables: 'T°', 'pH', 'Ce', 'TDS', 'OD', '%Sat. O', 'Ca', 'PO4', 'P T', 'Fe', 'NO3', 'NO2', 'NT', 'NKT', 'Si T', 'Turbidez', 'Crec_algal_Ausente', 'Crec_algal_Inicial', 'Crec_algal_Mediana', 'Crec_algal_Alta', 'Crec_algal_Muy Alta', '%Cob_algal_ausente', '%Cob_algal_Pequenas colonias', '%Cob_algal_Mediana', '%Cob_algal_Alta', '%Cob_algal_Muy Alta', 'Didymo'.

3.3.2. Modelado y entrenamiento

La parte dos se subdividió en tres partes, la primera en *clustering*, luego clasificación y finalmente, la realización de clasificación de nodos de un grafo.

3.3.2.1. *Clustering*

El objetivo de este procedimiento es realizar un análisis exhaustivo de datos utilizando técnicas de preprocesamiento, reducción de dimensionalidad y algoritmos de clustering. Se busca obtener una comprensión detallada de la estructura subyacente en los datos, identificar patrones significativos y asignar los registros a clústeres relevantes. A través de métodos como *K-Means, Gaussian Mixture Model* (GMM) y *Fuzzy C-Means* (FCM), se pretende explorar distintas perspectivas de agrupación en los datos, permitiendo una toma de decisiones informada basada en la naturaleza intrínseca de la información contenida en el dataset:

1. **Carga de Datos:**
   - Se cargaron los datos desde del archivo CSV “datanuevo.csv".

2. **Preprocesamiento y Estandarización:**
   - Se realizaron operaciones de preprocesamiento y estandarización de los datos utilizando la biblioteca *scikit-learn*.

3. **Análisis de Componentes Principales (PCA):**
   - Se aplicó el método de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. Se utilizó la visualización amarilla (`*yellowbrick*`) para visualizar la varianza explicada por cada componente principal.

4. **K-Means:**
   - Se realizó un análisis de *K-Means* para determinar el número óptimo de clústeres. Se utilizó el método del codo y el coeficiente *Silhouette* para esta determinación. Los datos se asignaron a clústeres utilizando *K-Means* con el número óptimo de clústeres.

5. **Gaussian Mixture Model (GMM):**
   - Se aplicó un modelo de mezcla gaussiana utilizando *GridSearchCV* para encontrar los mejores parámetros. Se evaluaron los resultados utilizando el coeficiente *Silhouette*.

6. **Fuzzy C-Means (FCM):**
   - Se aplicó el algoritmo *Fuzzy C-Means* para realizar *clustering* difuso. Se evaluaron los resultados utilizando el coeficiente *Silhouette*.

7. Visualización de Resultados:

- Se realizaron proyecciones en dos dimensiones utilizando PCA para visualizar los resultados de *K-Means* y FCM.

8. Resultados Finales:

- Se recopilaron y compararon los valores de coeficiente *Silhouette* de los diferentes modelos aplicados, incluyendo *K-Means*, GMM y FCM.

En resumen, el procedimiento abarcó desde la carga y estandarización de datos hasta la aplicación de diferentes técnicas de *clustering* y visualización para comprender la estructura subyacente de los datos.

3.3.2.2. Clasificación

El objetivo de este procedimiento es aplicar un modelo de clasificación *Random Forest* y redes neuronales para predecir las categorías asociadas a las variables "Crec_algal_Ausente," "Crec_algal_Inicial," "Crec_algal_Mediana," "Crec_algal_Alta," y "Crec_algal_Muy Alta" en función de otras características del conjunto de datos.

1. Carga de Datos:

- Se cargaron los datos desde el archivo "Dataset_analisis.csv" utilizando la biblioteca Pandas.

2. Eliminación de variables binarias

- Se eliminaron las variables binarias que no aportaban información al modelo

3. Definición de Características y Variable Objetivo:

- Se definieron las características (`X`) excluyendo las columnas "rio" y las etiquetas de crecimiento algal. La variable objetivo (`y`) se definió como las columnas de crecimiento algal.

4. División del Conjunto de Datos:

- El conjunto de datos se dividió en conjuntos de entrenamiento *(`X_train`, `y_train`)* y prueba (*`X_test`, `y_test`*) utilizando la función *`train_test_split`* de *scikit-learn*.

5. Inicialización del Modelo *Random Forest*:

- Se inicializó el modelo de clasificación *Random* Forest con 100 estimadores y una semilla aleatoria de 42.

6. Ajuste del Modelo:

- El modelo *Random* Forest se ajustó al conjunto de entrenamiento utilizando la función `*fit*`.

7. Evaluación del Rendimiento del Modelo:

- Se calcularon métricas de rendimiento, incluyendo la precisión global (`*accuracy*`) y métricas específicas para cada categoría utilizando la función `*classification_report*`. Estas métricas fueron impresas para proporcionar una evaluación detallada del rendimiento del modelo en el conjunto de prueba.

8. Definición red neuronal

- Se definió una red neuronal con 13 capas, 1 capa de entrada 1 de salida y 12 intermedias. Las capas intermedias tienen 100, 90, 80, 70, 60, 50, 40, 30, 20, 10 y 10 neuronas cada una, y la capa de salida tiene 5 neuronas (1 para cada resultado de clasificación) (Figura 20).

9. Entrenamiento y validación del modelo

- Se entreno el modelo con 200 épocas y se obtuvieron las métricas de validación.

En resumen, este procedimiento permitió entrenar un modelo *Random Forest* y de red neuronal para evaluar su rendimiento en la predicción de las categorías de crecimiento algal.

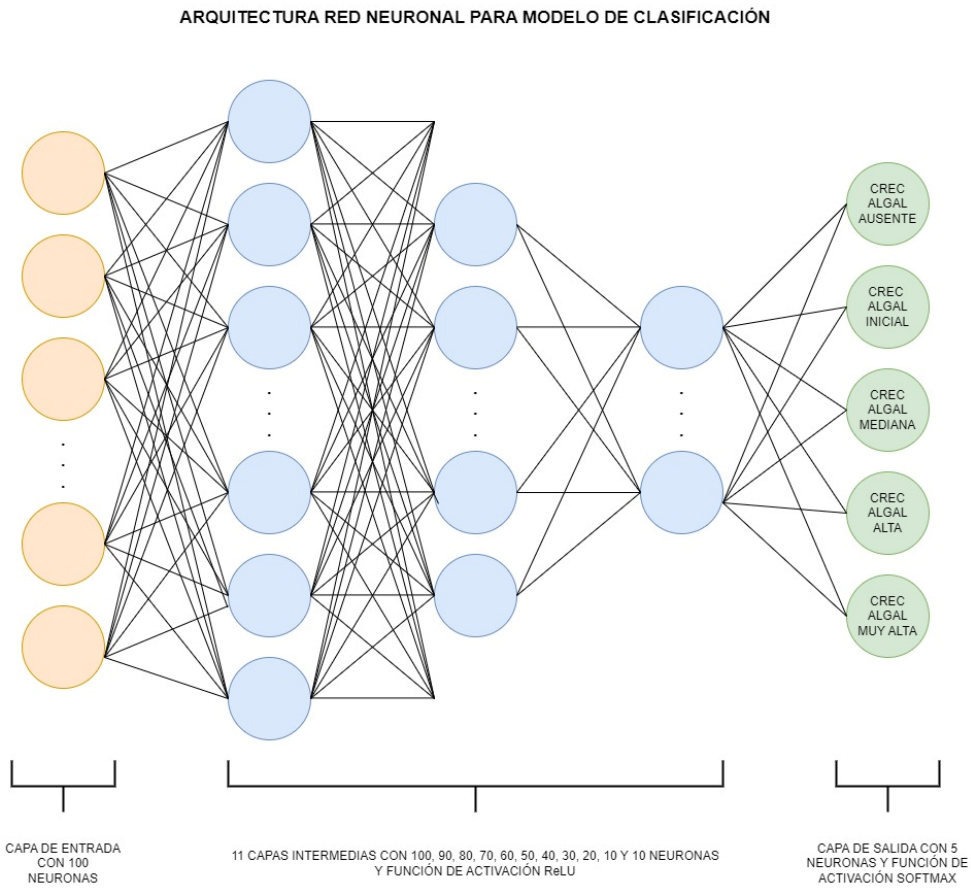


Figura 20: Red neuronal para modelo de clasificación. Fuente: Elaboración propia

3.3.2.3. Grafos

Procedimiento para la conversión de un archivo .csv en grafo y posterior clasificación de nodos:

1. **Carga de datos**
   - El archivo que se obtuvo del EDA en formato .csv fue cargado al nuevo notebook con el delimitador “;”.
   - Se verificó los nombres de las columnas a trabajar.

2. **Definición de variables**
   - Se definió la variable “atributos” con todos los atributos que tendrá cada nodo del grafo.
   - Se definió la variable “etiquetas” con la columna creada en el paso anterior para definir las clases del modelo.

3. **Creación de elementos para el grafo**
    - Se creó la matriz de atributos en formato tensor con todos los atributos definidos.
    - Se creó el vector formato tensor con etiquetas para el modelo de clasificación.
    - Se crearon las conexiones en formato tensor para el grafo, combinando todos los nodos entre sí. Cada nodo corresponde a una instancia.
    - Se crearon conexiones específicas, ya que en el punto anterior, todos los nodos se relacionaban entre sí.

4. **Creación del grafo**
    - Se creó un cargador de dato en *batch*, permitirá mayor velocidad en fases posteriores.
    - Se armó el grafo con los elementos del paso 4.
    - Se estableció el tipo de dato de los elementos.
    - Se verifica la calidad del grafo.

5. **Datos de entrenamiento, validación y prueba**
    - Se dividió el grafo para obtener datos de entrenamiento (70% de los datos), validación (20% de los datos) y prueba (10% de los datos).

6. **Entrenamiento y validación del modelo**
    - Se crearon 2 funciones, la primera para entrenar el modelo de red de neuronal de grafos y la segunda para evaluar la precisión.
    - Para el entrenamiento, se ingresó el modelo, el grafo, un optimizador, criterio y número de épocas.
    - Para la validación, se ingresó el modelo, grafo y los datos de prueba.

7. **Red neuronal perceptrón multicapa**
    - Se diseñó una red neuronal (Figura 21) para la clasificación del grafo, con capas intermedias, funciones de activación y salida. Esto se hizo a través de una clase llamada “MLP” con dos funciones, la primera función con la definición de todas las capas (entrada, intermedias y salida), número de neuronas y funciones de activación y la segunda función define los datos que utilizará para el posterior entrenamiento, indicando solo la matriz de atributos.
    - Posterior a esto, se entrenó el modelo con el optimizador Adam, criterio *CrossEntropyLoss* y con una cierta cantidad de épocas.
    - Finalmente se validó con los datos de prueba, retornando la precisión del modelo.

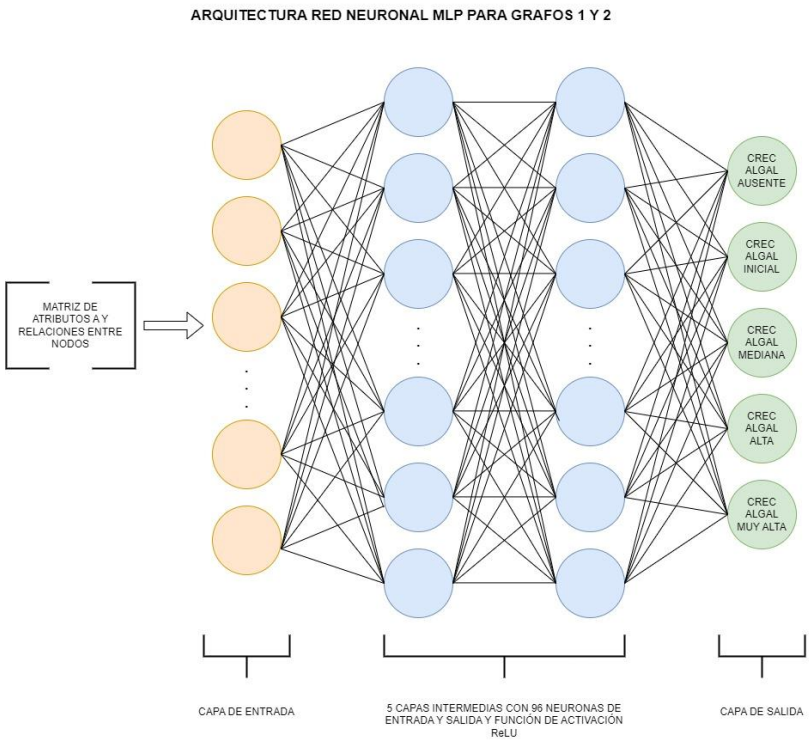


Figura 21: Arquitectura de red perceptrón neuronal (MLP). Fuente: Elaboración propia.

8. Red neuronal convolucional

- Se diseñó una red neuronal convolucional (Figura 22) para la clasificación de grafo, con capas intermedias, funciones de activación y salida. Esto se hizo a través de una clase llamada “GCN” con dos funciones, la primera función con la definición de cada capa y número de neuronas (entrada, intermedias, salida), y la segunda función define los datos que utilizará para el posterior entrenamiento y arma la red agregando las funciones de activación.
- Posterior a esto, se entrenó el modelo con el optimizador Adam, criterio *CrossEntropyLoss* y con una cierta cantidad de épocas.
- Finalmente se validó con los datos de prueba, retornando la precisión del modelo.

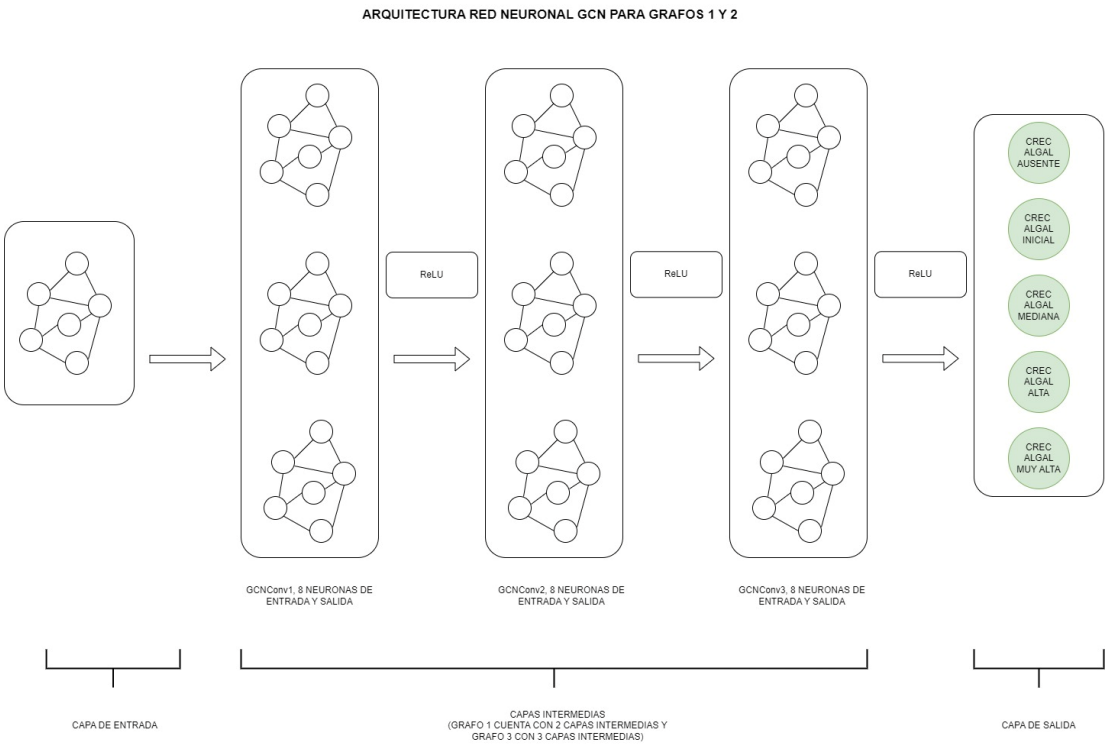


Figura 22: Arquitectura red neuronal convolucional. Fuente: Elaboración propia.

9. Red neuronal convolucional Normalizada

- Se diseñó una red neuronal convolucional normalizada (Figura 23) para la clasificación de grafo, con capas intermedias, funciones de activación y salida y además una capa de normalización de la matriz de atributos. Esto se hizo a través de una clase llamada “GCNNorm” con dos funciones, la primera función con la definición de cada capa y número de neuronas (entrada, normalización, intermedias, salida), y la segunda función define los datos que utilizará para el posterior entrenamiento y arma la red agregando las funciones de activación.
- Posterior a esto, se entrenó el modelo con el optimizador Adam, criterio *CrossEntropyLoss* y con una cierta cantidad de épocas.
- Finalmente se validó con los datos de prueba, retornando la precisión del modelo.

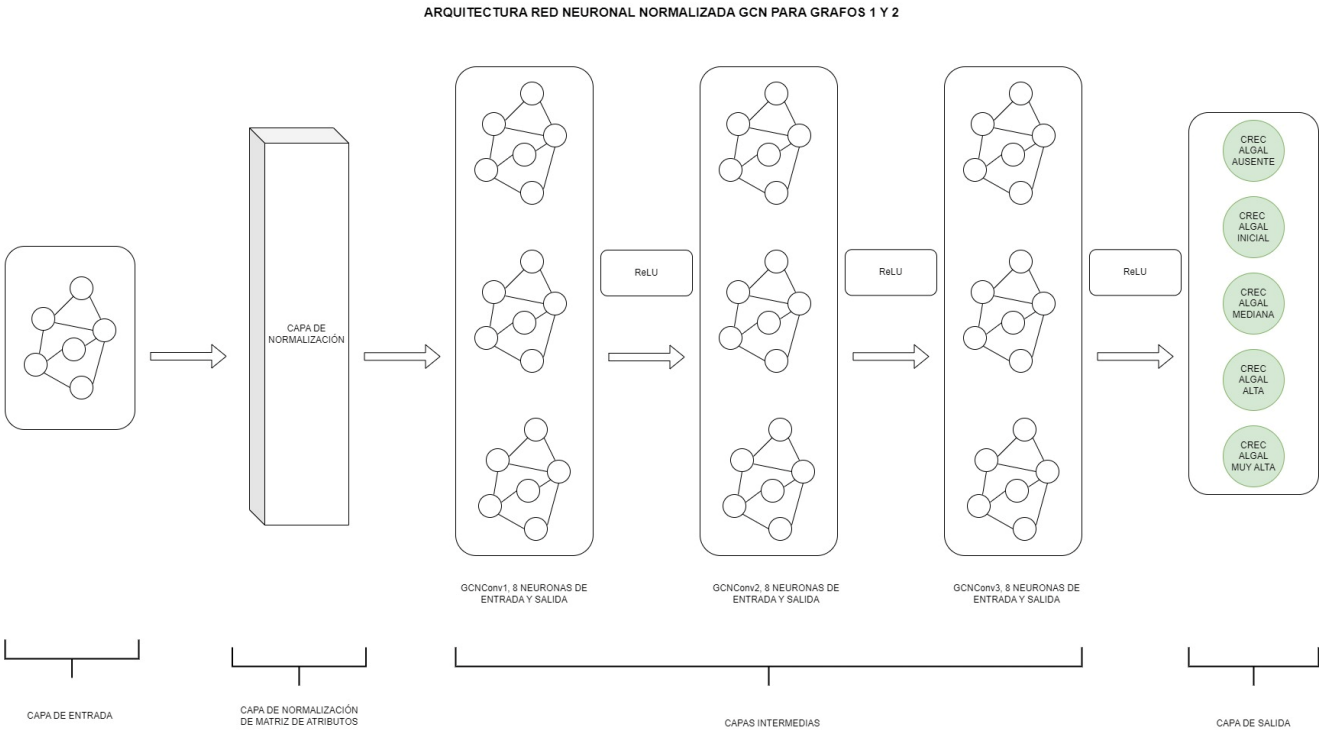


Figura 23: Arquitectura red neuronal convolucional normalizada. Fuente: Elaboración propia.

4. Marco Normativo

En el desarrollo de este trabajo de fin de master, se va a utilizar un conjunto de datos con información real, solicitada a la Subsecretaría de Pesca y Acuicultura a través de la consulta N° AH002T-0005906, la cual fue autorizada en virtud de lo establecido en el artículo 13° de la Ley N°20.285 sobre Acceso a la Información Pública, correspondiente al proyecto "Monitoreo, prospección e investigación de la especie plaga Didymosphenia geminata en ecosistemas fluviales y lacustres de la zona centro, sur y austral de Chile" que es ejecutado por el Departamento de Medio Ambiente de la División de Investigación en Acuicultura del Instituto de Fomento Pesquero (sede Puerto Montt), que es parte del programa de investigación permanente de la Subsecretaría de Pesca y Acuicultura, financiado por el Ministerio de Economía (Anexo II).

5. Desarrollo específico de la contribución

5.1. Set de datos

El set de datos utilizado consta de las siguientes características:

- 140 atributos
- 1605 instancias
- Tipo de datos: numérico, flotante, categórico
- Caracteres especiales: 's/m', 's/i', 's/n', 's/h', 'ds/m', '<'
- Valores nulos en varias columnas

Y contiene los siguientes atributos (Tabla *6*):

Tabla 6: Metadatos del set de datos.

N°	Atributo	Tipo	Descripción
1	ID	Numérica	Identificador de la muestra.
2	nom_est	Categórica	Nombre de la estación de muestreo.
3	cod_region	Numérica	Código región.
4	region	Categórica	Nombre región.
5	cod_provincia	Numérica	Código provincia.
6	nom_provincia	Categórica	Nombre provincia.
7	cod_comuna	Numérica	Código comuna.
8	comuna	Categórica	Nombre comuna
9	cod_cuenca	Numérica	código de la cuenca muestreada
10	cuenca	Categórica	Nombre de la cuenca muestreada.
11	cod_subcuenca	Numérica	código de la subcuenca muestreada
12	subcuenca	Categórica	Nombre de la subcuenca muestreada.
13	cod_sscuenca	Numérica	Código de la subsubcuenca muestreada
14	subsubcuenca	Categórica	Nombre de la subsubcuenca muestreada.
15	rio	Categórica	nombre del río muestreado
16	altitud	Numérica	Altitud del sitio de muestreo.
17	este	Numérica	Coordenadas del sitio de muestreo.
18	norte	Numérica	
19	huso	Numérica	Huso cartográfico.
20	Lat	Numérica	Coordenadas del sitio de muestreo.
21	Long	Numérica	
22	Nom_proy	Categórica	Nombre del Proyecto.
23	ejecutor	Categórica	Ejecutor del Proyecto.
24	periodo_inicio	Fecha	Fecha inicio de la etapa del proyecto.

N°	Atributo	Tipo	Descripción
25	periodo_final	Fecha	Fecha término de la etapa del proyecto.
26	Etapa	Categórica	Número de Etapa
27	Camp	Categórica	Campaña
28	Cod_est	Numérica	Código estación de muestreo.
29	epoca_muest	Categórica	época en que se hizo el muestreo
30	fecha	Fecha	Fecha de la toma de muestra.
31	Hora	Hora	Hora de la toma de muestra.
32	cod_lab	Numérica	Código de la muestra analizada en laboratorio.
33	T°	Numérica	Medición de temperatura del agua
34	pH	Numérica	Medición de pH. Escala de pH
35	Ce	Numérica	Medición de conductividad eléctrica. uS/cm2
36	TDS	Numérica	Medición de sólidos totales disueltos en agua. mg/L
37	OD	Numérica	Medición de oxígeno disuelto en agua. mg/L
38	%Sat. O	Numérica	Saturación de oxígeno disuelto en agua. %
39	Ca	Numérica	Medición de calcio disuelto en agua. mg/L
40	PO4	Numérica	Medición de fosfato disuelto en agua. mg/L
41	P T	Numérica	Medición de fosforo total disuelto en agua. mg/L
42	Fe	Numérica	Medición de hierro disuelto en agua. mg/L
43	NO3	Numérica	Medición de nitrato disuelto en agua. mg/L
44	NO2	Numérica	Medición de nitrito disuelto en agua. mg/L
45	NT	Numérica	Medición de nitrógeno total disuelto en agua. mg/L
46	NKT	Numérica	Medición de nitrógeno Kjeldahl disuelto en agua. mg/L
47	Si T	Numérica	Medición de silica total disuelto en agua. mg/L
48	Turbidez	Numérica	Medición de turbidez en agua. NTU
49	prof_f	Numérica	Profundidad del fondo del río en metros
50	vel_0.2_f	Numérica	Velocidad en m/s a 0.2, 0.6 y 0.8 metros de profundidad
51	vel_0.6_f	Numérica	
52	vel_0.8_f	Numérica	
53	prof_m	Numérica	Velocidad media en m/s
54	vel_0.2_m	Numérica	Velocidad media a 0.2, 0.6 y 0.8 metros de profundidad en m/s
55	vel_0.6_m	Numérica	
56	vel_0.8_m	Numérica	
57	Sust_fondo_Roca madre	Numérica	Presencia (1) o ausencia (0) de sustrato Indicado en nombre del atributo
58	Sust_fondo_Bloques y piedras	Numérica	
59	Sust_fondo_Bolones	Numérica	

N°	Atributo	Tipo	Descripción
60	Sust_fondo_Gravas	Numérica	Presencia (1) o ausencia (0) de sustrato Indicado en nombre del atributo
61	Sust_fondo_Arena	Numérica	
62	Sust_fondo_Fango	Numérica	
63	Sust_fondo_Macrofitas+algas filamentosas	Numérica	
64	Sust_fondo_Grandes residuos lenosos	Numérica	
65	Sust_fondo_Camadas de hojas	Numérica	
66	Sust_fondo_Sustratos artificiales	Numérica	
67	Sust_fondo_SEDIMENTARIA	Numérica	
68	Diam_medio_bolones	Numérica	Diámetro medio de bolones en centímetros
69	Diam_max_bolones	Numérica	Diámetro máximo de bolones en centímetros
70	%Enfangam_bolones	Numérica	Porcentaje de enfangamiento de bolones. %
71	Ac_entorno_Agricola	Numérica	Entorno del sitio de muestreo. Variable binaria, el valor 1 nombre del atributo
72	Ac_entorno_Ganadero	Numérica	
73	Ac_entorno_Forestal	Numérica	
74	Ac_entorno_Balseo	Numérica	
75	Ac_entorno_Poblado	Numérica	
76	Ac_entorno_Infraestructura_vial	Numérica	
77	Ac_entorno_Desagüe	Numérica	
78	Ac_entorno_Basural	Numérica	
79	Ac_entorno_Embalse	Numérica	
80	Ac_entorno_Extraccion_de_aridos	Numérica	
81	Ac_entorno_Area_protegida	Numérica	
82	Ac_entorno_Toma_de_agua	Numérica	
83	Ac_entorno_Pesca_deportiva	Numérica	
84	Ac_entorno_Deportes_acuaticos	Numérica	
85	Ac_entorno_Camping-picnic	Numérica	
86	Ac_entorno_Embarcadero	Numérica	
87	Ac_entorno_Piscicultura	Numérica	
88	Ac_entorno_Estacion_DGA	Numérica	
89	Ac_entorno_Balneario	Numérica	
90	Ac_entorno_area_de_pesca	Numérica	
91	Ac_entorno_Cabanas	Numérica	
92	Ac_entorno_Planta_de_tratamiento	Numérica	
93	Ac_entorno_Canalizacion	Numérica	
94	Ac_entorno_Acceso_publico	Numérica	
95	Cond_met_lluvioso	Numérica	Condición meteorológica lluviosa el día de la toma de muestra
96	%CobNubes	Numérica	Porcentaje de cobertura de las nubes día muestreado
97	T_tramo_Duna y estria	Numérica	Tipo de hábitat del sitio de muestreo. 1 tipo descrito en el nombre de la columna, 0 no pertenece a ese tipo
98	T_tramo_Rabiones y poza	Numérica	Tipo de hábitat del sitio de muestreo. 1 tipo descrito en el nombre de la columna, 0 no pertenece a ese tipo
99	T_tramo_Lechos planos	Numérica	
100	T_tramo_Gradas y pozas	Numérica	
101	T_tramo_Cascadas	Numérica	

N°	Atributo	Tipo	Descripción
102	T_tramo_Lago	Numérica	Tipo de hábitat del sitio de muestreo. 1 tipo descrito en el nombre de la columna, 0 no pertenece a ese tipo
103	T_tramo_Desague Lago	Numérica	
104	T_tramo_Canal Lateral	Numérica	
105	T_habitat_R. profundo	Numérica	
106	T_habitat_R. somero	Numérica	
107	T_habitat_L. profundo	Numérica	
108	T_habitat_L. somero	Numérica	
109	T_habitat_Poza	Numérica	
110	T_habitat_Lago	Numérica	
111	Clar_agua_Clara	Numérica	Claridad del agua del sitio de muestreo
112	Clar_agua_Ligeramente Turbia	Numérica	
113	Clar_agua_Turbia	Numérica	
114	Clar_agua_Muy turbia	Numérica	
115	%sombra_tramo_Sombreado con ventana	Numérica	% de sombra en el sitio de muestreo
116	%sombra_tramo_Sombreado total	Numérica	
117	%sombra_tramo_Grandes claros	Numérica	
118	%sombra_tramo_Expuestos	Numérica	
119	Form_canal_Serpenteante	Numérica	Forma del sitio de muestreo. 1 forma descrita en el nombre del atributo, 0 no corresponde a esa forma
120	Form_canal_Sinuoso	Numérica	
121	Form_canal_Trenzado	Numérica	
122	Form_canal_Encajonado	Numérica	
123	Form_canal_Con alteracion_de_cauce	Numérica	
124	Form_canal_Recto	Numérica	
125	Form_canal_Lago	Numérica	
126	Form_canal_Canales laterales	Numérica	
127	Crec_algal_Ausente	Numérica	Crecimiento del alga en el sitio de muestreo. 1 crecimiento descrito en el nombre del atributo, 0 no es ese crecimiento
128	Crec_algal_Inicial	Numérica	
129	Crec_algal_Mediana	Numérica	
130	Crec_algal_Alta	Numérica	
131	Crec_algal_Muy Alta	Numérica	
132	%Cob_algal_ausente	Numérica	Porcentaje de cobertura del alga en el sitio de muestreo. 1 cobertura descrita en el nombre del atributo, 0 no es esa cobertura
133	%Cob_algal_Pequenas colonias	Numérica	
134	%Cob_algal_Mediana	Numérica	
135	%Cob_algal_Alta	Numérica	
136	%Cob_algal_Muy Alta	Numérica	
137	Cond_cpo_agua_Normal	Numérica	Condición del cuerpo de agua en el sitio de muestreo. 1 condición descrita en el nombre del atributo, 0 no es esa condición
138	Cond_cpo_agua_Espumas superficie	Numérica	
139	Cond_cpo_agua_Descargas	Numérica	
140	Cond_cpo_agua_Material Aloctono	Numérica	
141	Didymo	Numérica	Concentración de células de didymo células/cm2

Fuente: Datos obtenidos a través de la Ley de transparencia de la Subsecretaría de Pesca y Agricultura, solicitud de Información AH002T0005906

Luego de la limpieza de datos se obtuvo un set de datos completo, sin caracteres especiales o espacios en blanco, y solo se almacenaron las variables a utilizar en etapas posteriores, tal como se muestra en la Figura 24.

a)

	T°	pH	Ce	TDS	OD	%Sat. O	Ca	PO4	P T	Fe	NO3	NO2	NT	NKT	Si T	Turbidez	Crec_algal_Ausente	%Cob_algal_ausente	Didymo
0	14.8	8.2	80	40	11.8	118.4	7,9	<1,0	0,6	s/i	<0,20	<0,10	2,9	2,9	1,0	0,7	0	0	0
1	15.6	6.7	73	36	12.2	124.2	9,2	<1,0	0,7	s/i	<0,20	<0,10	4,6	4,6	21,5	0,2	0	0	0
2	11.1	7.1	152	76	s/m	s/m	5,5	<1,0	<0,20	s/i	<0,20	<0,10	1,9	1,9	22,7	<0,20	0	0	0
3	15.2	7.8	82	13.8	11.8	111.8	4,2	<1,0	1,0	s/i	<0,20	<0,10	2,0	2,0	14,9	0,3	0	0	0
4	13.9	7.4	31	15	10.4	102.3	4,5	<1,0	<0,20	s/i	<0,20	<0,10	2,0	2,0	9,0	0,6	0	0	0
5	9.9	7.7	53	26	7.8	103.8	3,4	<1,0	0,6	s/i	<0,20	<0,10	2,6	2,6	14,3	0,4	0	0	0
6	10.8	7	85	43	7.23	s/m	7,0	<1,0	0,9	s/i	<0,20	<0,10	1,5	1,5	16,7	0,4	0	0	0
7	9.6	7.2	89	45	9.9	94.3	2,6	<1,0	<0,20	s/i	<0,20	<0,10	1,9	1,9	6,9	0,4	0	0	0
8	9.9	7.3	88	46	11.6	103.6	2,1	<1,0	<0,20	s/i	<0,20	<0,10	2,7	2,7	4,9	0,4	0	0	0
9	12.1	7.3	87	82	12.3	120.2	3,2	<1,0	0,7	s/i	<0,20	<0,10	1,6	1,6	12,8	0,5	0	0	0

b)

	T°	pH	Ce	TDS	OD	%Sat. O	Ca	PO4	P T	Fe	NO3	NO2	NT	NKT	Si T	Turbidez	Crec_algal_Ausente	%Cob_algal_ausente	Didymo
0	14.8	8.2	80.0	40.0	11.800000	118.400000	7.9	1.0	0.6	0.021000	0.2	0.1	2.9	2.9	1.0	0.7	0.0	0.0	0.0
1	15.6	6.7	73.0	36.0	12.200000	124.200000	9.2	1.0	0.7	0.021000	0.2	0.1	4.6	4.6	21.5	0.2	0.0	0.0	0.0
2	11.1	7.1	152.0	76.0	10.035000	101.520000	5.5	1.0	0.2	0.021000	0.2	0.1	1.9	1.9	22.7	0.2	0.0	0.0	0.0
3	15.2	7.8	82.0	13.8	11.800000	111.800000	4.2	1.0	1.0	0.021000	0.2	0.1	2.0	2.0	14.9	0.3	0.0	0.0	0.0
4	13.9	7.4	31.0	15.0	10.400000	102.300000	4.5	1.0	0.2	0.021000	0.2	0.1	2.0	2.0	9.0	0.6	0.0	0.0	0.0
5	9.9	7.7	53.0	26.0	7.800000	103.800000	3.4	1.0	0.6	0.021000	0.2	0.1	2.6	2.6	14.3	0.4	0.0	0.0	0.0
6	10.8	7.0	85.0	43.0	7.230000	105.770000	7.0	1.0	0.9	0.021000	0.2	0.1	1.5	1.5	16.7	0.4	0.0	0.0	0.0
7	9.6	7.2	89.0	45.0	9.900000	94.300000	2.6	1.0	0.2	0.023152	0.2	0.1	1.9	1.9	6.9	0.4	0.0	0.0	0.0
8	9.9	7.3	88.0	46.0	11.600000	103.600000	2.1	1.0	0.2	0.023152	0.2	0.1	2.7	2.7	4.9	0.4	0.0	0.0	0.0
9	12.1	7.3	87.0	82.0	12.300000	120.200000	3.2	1.0	0.7	0.030000	0.2	0.1	1.6	1.6	12.8	0.5	0.0	0.0	0.0

c)

	T°	pH	Ce	TDS	OD	%Sat. O	Ca	PO4	P T	Fe	NO3	NO2	NT	NKT	Si T	Turbidez	Crec_algal_Ausente	%Cob_algal_ausente	Didymo
0	14.8	8.2	80.0	40.0	11.8	118.4	7.9	1.0	0.6	0.021	0.2	0.1	2.9	2.9	1.0	0.7	0.0	0.0	0.0
1	15.6	6.7	73.0	36.0	12.2	124.2	9.2	1.0	0.7	0.021	0.2	0.1	4.6	4.6	21.5	0.2	0.0	0.0	0.0
2	11.1	7.1	152.0	76.0	10.0	101.5	5.5	1.0	0.2	0.021	0.2	0.1	1.9	1.9	22.7	0.2	0.0	0.0	0.0
3	15.2	7.8	82.0	13.8	11.8	111.8	4.2	1.0	1.0	0.021	0.2	0.1	2.0	2.0	14.9	0.3	0.0	0.0	0.0
4	13.9	7.4	31.0	15.0	10.4	102.3	4.5	1.0	0.2	0.021	0.2	0.1	2.0	2.0	9.0	0.6	0.0	0.0	0.0
5	9.9	7.7	53.0	26.0	7.8	103.8	3.4	1.0	0.6	0.021	0.2	0.1	2.6	2.6	14.3	0.4	0.0	0.0	0.0
6	10.8	7.0	85.0	43.0	7.2	105.8	7.0	1.0	0.9	0.021	0.2	0.1	1.5	1.5	16.7	0.4	0.0	0.0	0.0
7	9.6	7.2	89.0	45.0	9.9	94.3	2.6	1.0	0.2	0.023	0.2	0.1	1.9	1.9	6.9	0.4	0.0	0.0	0.0
8	9.9	7.3	88.0	46.0	11.6	103.6	2.1	1.0	0.2	0.023	0.2	0.1	2.7	2.7	4.9	0.4	0.0	0.0	0.0
9	12.1	7.3	87.0	82.0	12.3	120.2	3.2	1.0	0.7	0.030	0.2	0.1	1.6	1.6	12.8	0.5	0.0	0.0	0.0

d)

	T°	categoria	pH	Ce	TDS	OD	%Sat. O	Ca	PO4	P T	Fe	NO3	NO2	NT	NKT	Si T	Turbidez	Crec_algal_Ausente	%Cob_algal_ausente	Didymo
0	14.8	%Cob_algal_Alta	8.2	80.0	40.0	11.8	118.4	7.9	1.0	0.6	0.021	0.2	0.1	2.9	2.9	1.0	0.7	0.0	0.0	0.0
1	15.6	%Cob_algal_Alta	6.7	73.0	36.0	12.2	124.2	9.2	1.0	0.7	0.021	0.2	0.1	4.6	4.6	21.5	0.2	0.0	0.0	0.0
2	11.1	%Cob_algal_Alta	7.1	152.0	76.0	10.0	101.5	5.5	1.0	0.2	0.021	0.2	0.1	1.9	1.9	22.7	0.2	0.0	0.0	0.0
3	15.2	%Cob_algal_Alta	7.8	82.0	13.8	11.8	111.8	4.2	1.0	1.0	0.021	0.2	0.1	2.0	2.0	14.9	0.3	0.0	0.0	0.0
4	13.9	%Cob_algal_Muy Alta	7.4	31.0	15.0	10.4	102.3	4.5	1.0	0.2	0.021	0.2	0.1	2.0	2.0	9.0	0.6	0.0	0.0	0.0
5	9.9	%Cob_algal_Alta	7.7	53.0	26.0	7.8	103.8	3.4	1.0	0.6	0.021	0.2	0.1	2.6	2.6	14.3	0.4	0.0	0.0	0.0
6	10.8	%Cob_algal_Alta	7.0	85.0	43.0	7.2	105.8	7.0	1.0	0.9	0.021	0.2	0.1	1.5	1.5	16.7	0.4	0.0	0.0	0.0
7	9.6	%Cob_algal_Mediana	7.2	89.0	45.0	9.9	94.3	2.6	1.0	0.2	0.023	0.2	0.1	1.9	1.9	6.9	0.4	0.0	0.0	0.0
8	9.9	%Cob_algal_Alta	7.3	88.0	46.0	11.6	103.6	2.1	1.0	0.2	0.023	0.2	0.1	2.7	2.7	4.9	0.4	0.0	0.0	0.0
9	12.1	%Cob_algal_Alta	7.3	87.0	82.0	12.3	120.2	3.2	1.0	0.7	0.030	0.2	0.1	1.6	1.6	12.8	0.5	0.0	0.0	0.0

Figura 24: Set de datos, se muestran solo las variables con las que se trabajan posteriormente y de las variables objetivo, solo se observan Crec_algal_Ausente y %Cob_algal_ausente. a) set de datos original, se observan caracteres especiales y datos faltantes (s/i, s/m). b) Set de datos tratados, se eliminaron caracteres especiales y se completaron los datos faltantes. c) set de datos truncado de acuerdo con las especificaciones entregadas por centro de investigación. d) set de fatos final, con la variable categoría. Fuente: Elaboración propia.

Los diferentes métodos se entrenaron con los siguientes atributos, con el objetivo de obtener resultados comparables:

- *Clustering:* 'T°', 'pH', 'Ce', 'TDS', 'OD', '%Sat. O', 'Ca', 'PO4', 'P T', 'Fe', 'NO3', 'NO2', 'NT', 'NKT', 'Si T', 'Turbidez', 'Didymo'

- *Ensembled learning* y red convolucional: 'T°', 'pH', 'Ce', 'TDS', 'OD', '%Sat. O', 'Ca', 'PO4', 'P T', 'Fe', 'NO3', 'NO2', 'NT', 'NKT', 'Si T', 'Turbidez', 'Didymo', 'Crec_algal_Ausente', 'Crec_algal_Inicial', 'Crec_algal_Mediana', 'Crec_algal_Alta', 'Crec_algal_Muy Alta', '%Cob_algal_Pequenas colonias', '%Cob_algal_Mediana', '%Cob_algal_Alta', '%Cob_algal_Muy Alta'

5.2. Análisis de Resultados

5.2.1. Métodos de Clasificación para el Crecimiento de Didymo en Sistemas Hídricos de Chile mediante Técnicas de Clustering

Con el objetivo de implementar un modelo de predicción del crecimiento de Didymo en los sistemas hídricos de Chile, se emplearon técnicas aprendizaje no supervisado de clustering utilizando la información proporcionada por el Instituto de Fomento Pesquero de Puerto Montt. En este contexto, se evaluaron tres métodos de *clustering: K-Means*, Gaussian *Mixtures* y *Fuzzy C-Means*, con el fin de identificar patrones y estructuras subyacentes en los datos.

Para todos los métodos se establecieron 5 *clusters*, ya que, en el set de datos original, 5 son las etiquetas de interés: Crecimiento algal Ausente, Crecimiento algal Inicial, Crecimiento algal Mediana, Crecimiento algal Alta y Crecimiento algal Muy Alta. También, se realizó con y sin la técnica PCA (reducción de dimensiones) debido a la cantidad de atributos (17 atributos)

K-Means:
En primer lugar, se analizó el método del codo, con el objetivo de establecer el número óptimo de *clusters*:

Cómo se observa en la Figura 25, no es posible determinar el número óptimo de *clusters* para el método K-Means con y sin aplicar la técnica PCA, no se observa un punto de inflexión prominente, por lo que se trabajó con “5”, como se explicó anteriormente, debido al número de etiquetas objetivo.

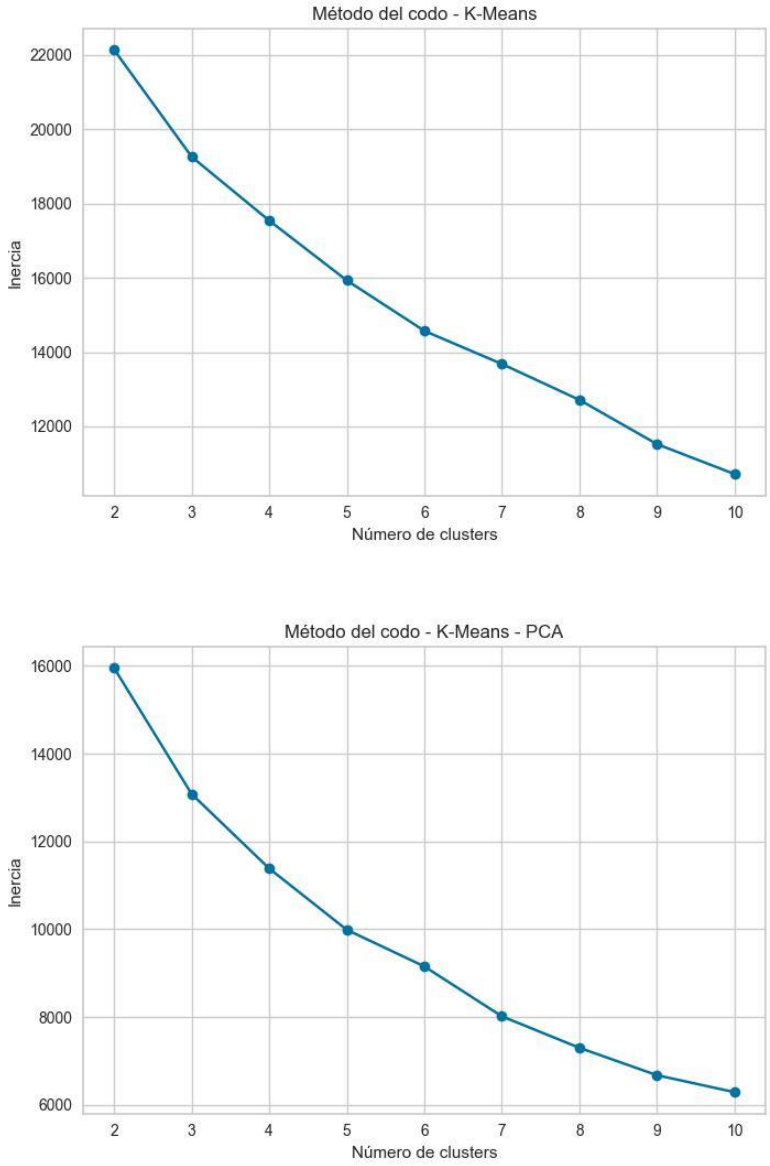


Figura 25: Curva método del codo para establecer número óptimo de *clusters* con el método K-Means con y sin la aplicación de técnica PCA. Fuente: Elaboración propia.

Al calcular el coeficiente *Silhouette*, se observa en la Figura *26* que el número óptimo es 3 o 5 *clusters*, por lo que se eligió trabajar con 5 *clusters*.

En la Tabla 7 se observa la distribución de las instancias en los distintos *clusters*. Alrededor del 60% de instancias se aglomeran en un solo *cluster*, y algunos *clusters* solo tienen 1, 2 o 3 instancias.

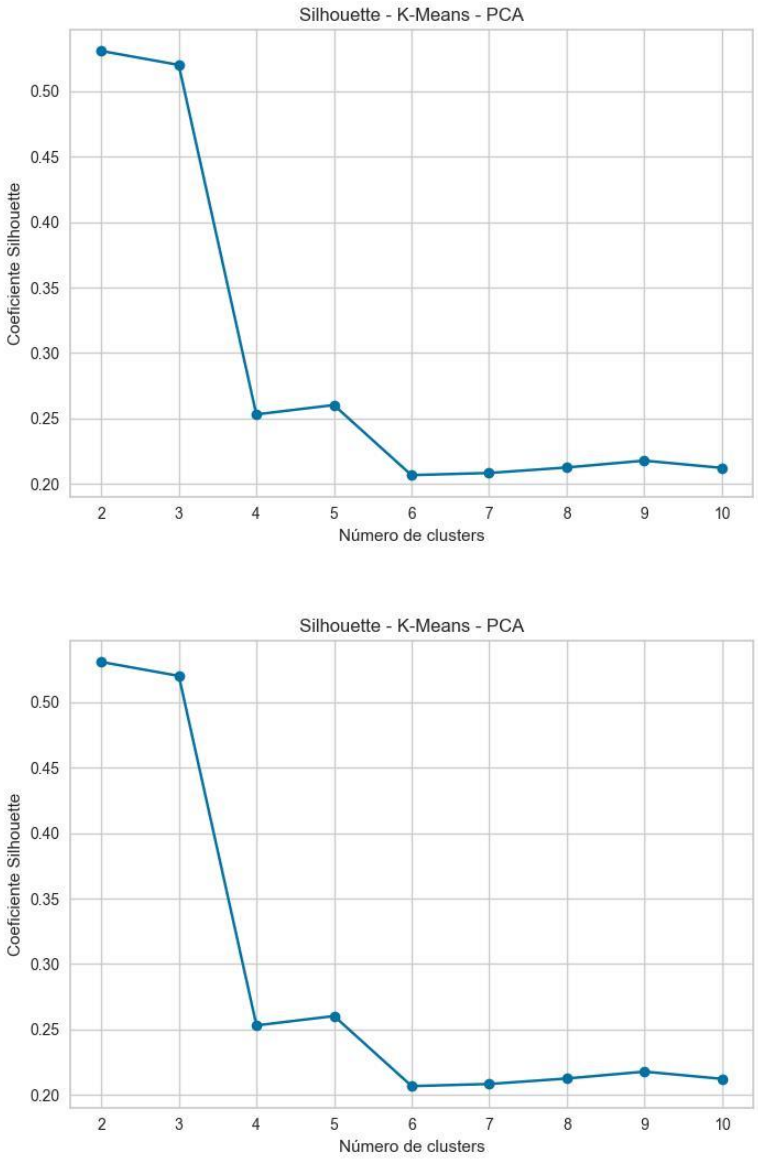


Figura 26: Coeficiente *silhouette* método K-Means con y sin aplicación de la técnica PCA. Fuente: Elaboración propia.

Tabla 7: Número de instancias asignadas en cada *cluster* con el método K-Means

Número de *clusters*	*K-Means*	*K-Means - PCA*
0	980	974
1	194	434
2	426	194
3	3	2
4	2	1

Fuente: Elaboración propia.

En paralelo, se calculó el coeficiente *Silhouette* para cada método:

- *K-Means*: 0.2788
- *K-Means – PCA*: 0.2895

Y se graficaron en 2 dimensiones los *clusters* (Figura 27) para un mejor entendimiento de la distribución de estos:

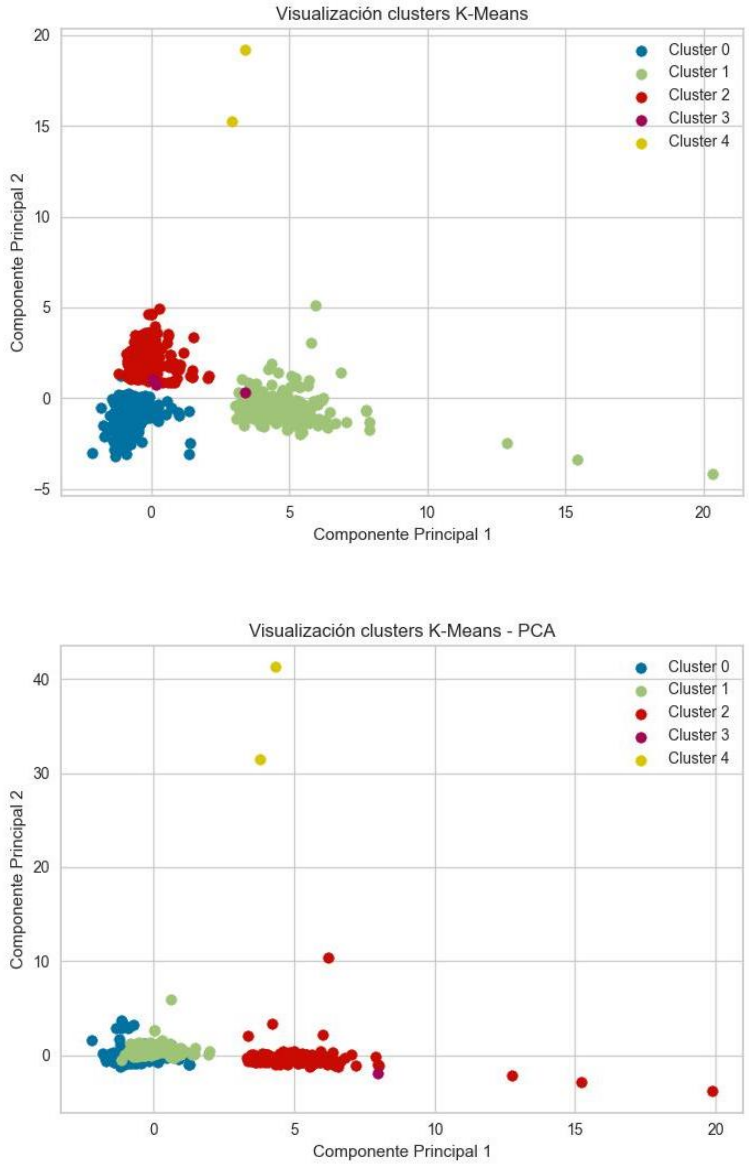


Figura 27: Visualización de *clusters* con método K-*Means* con y sin aplicar técnica PCA. Fuente: Elaboración propia.

En la técnica de *clustering* K-Means no se obtuvieron buenos resultados. Al aplicar la técnica PCA, no se observó una mejora significativa. Si bien se observaron *clusters* relativamente compactos y definidos, la asignación de instancias no es acorde a lo esperado.

Gaussian Mixtures:

Para la técnica de *clustering* a través del método *Gaussian Mixture*, se analizó en primer lugar el número óptimo de *clusters*, también aplicando la técnica PCA:

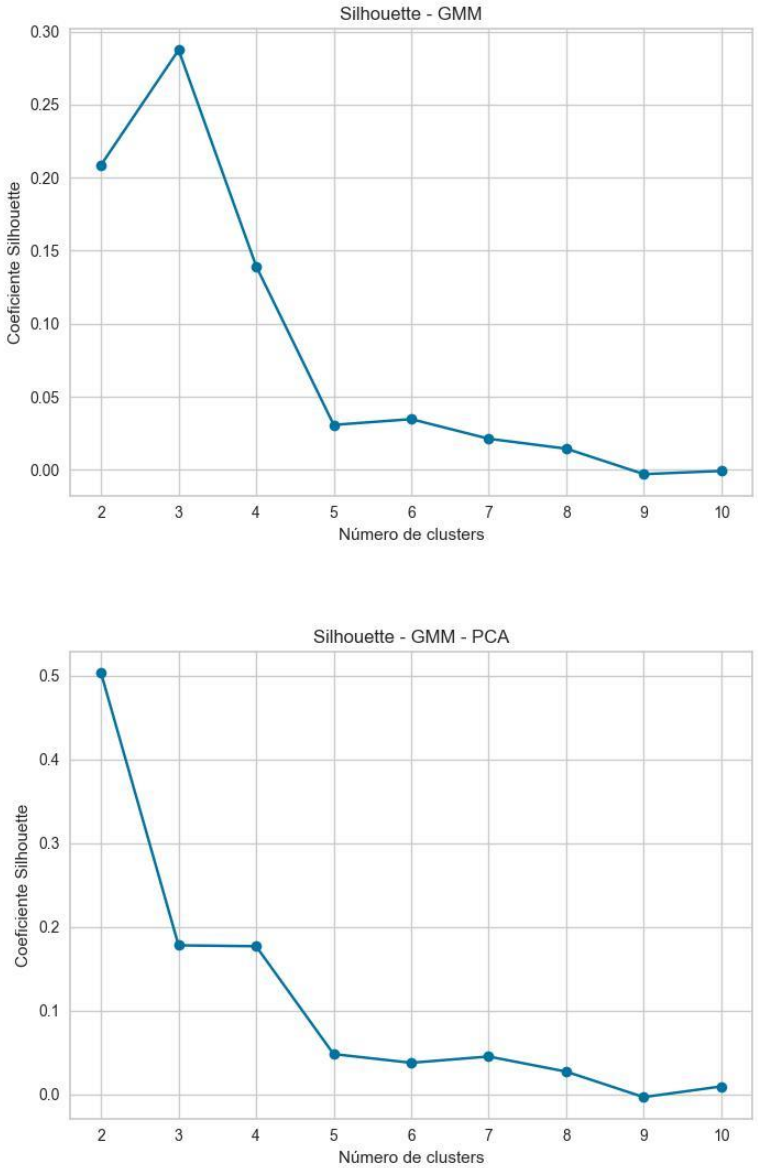


Figura 28: Coeficiente *silhouette* método *Gaussian Mixture* con y sin aplicación de la técnica PCA. Fuente: Elaboración propia.

Como se observó en la Figura *28*, para el método sin PCA, el número de *cluster* óptimos es 3 y para el método con PCA, el número óptimo es 4. Esto se contradice a lo esperado, que es 5 *clusters*.

Además, al analizar el número de instancias a cada *cluster*, se observó lo siguiente (Tabla 8):

Tabla 8: Número de instancias asignadas en cada *cluster* con el método *Gaussian Mixture*

Número de *cluster*	***Gaussian Mixture***	***Gaussian Mixture - PCA***
0	264	240
1	180	184
2	347	779
3	780	368
4	54	34

Fuente: Elaboración propia.

Si bien, la asignación entre ambos métodos es similar, alrededor de un 48% de las instancias son asignadas a un solo *cluster*.

En cuanto al coeficiente *silhouette*, se obtuvo lo siguiente:

- *Gaussian Mixture*: 0.1777
- *Gaussian Mixture* – PCA: 0.1133

Los *clusters* se graficaron en 2 dimensiones (Figura 29) para un mejor entendimiento de la distribución de estos.

Ambos coeficientes son similares, por lo que aplicar la técnica PCA no entrega una mejora significativa. Y gráficamente, no se observan *clusters* compactos y definidos (Figura 29).

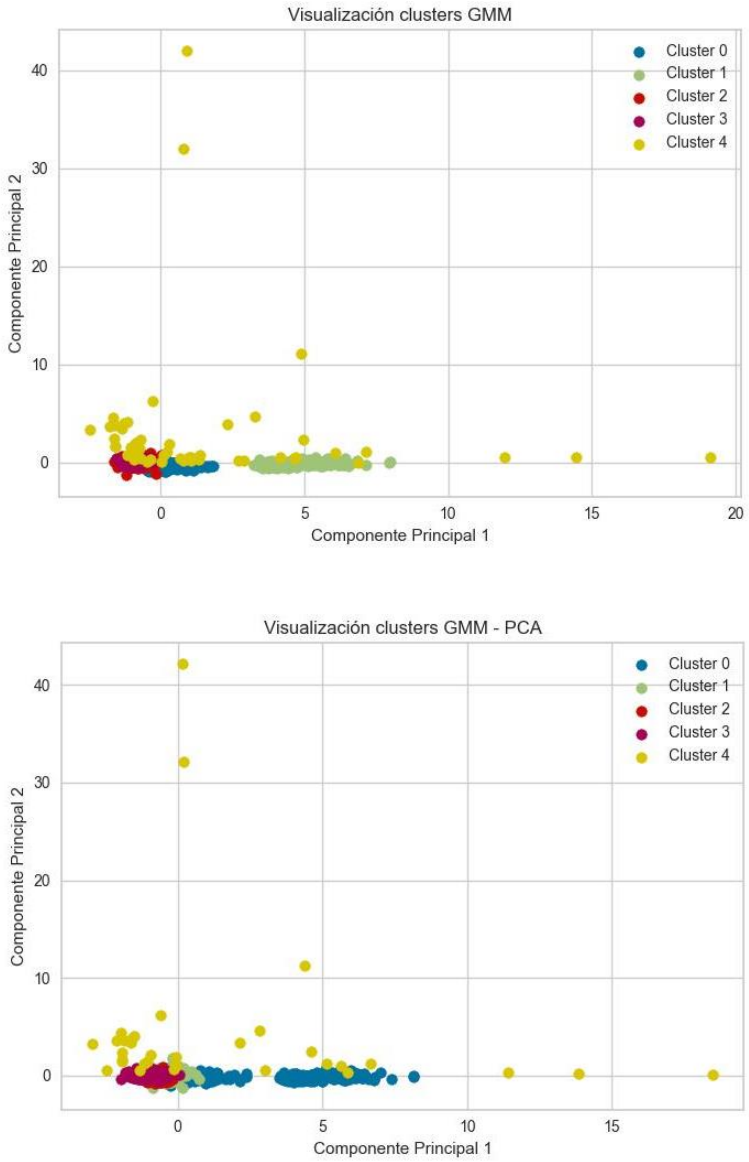


Figura 29: Visualización de *clusters* con método Gaussian Mixture con y sin aplicar técnica PCA.
Fuente: Elaboración propia

Fuzzy C-Means:

La tercera técnica de *clustering* aplicada fue *Fuzzy C-Means*, obteniéndose los siguientes resultados:

En la Figura 30, se observa que el número de *clusters* óptimo para *Fuzzy C-Means* es entre 2 y 7 *clusters*. Para el método con la técnica PCA, es difícil saber cuánto es el óptimo ya que la curva decrece conforme aumenta el número de *clusters*. Para ambos, se trabajará con 5 *clusters*, al igual que los métodos anteriores.

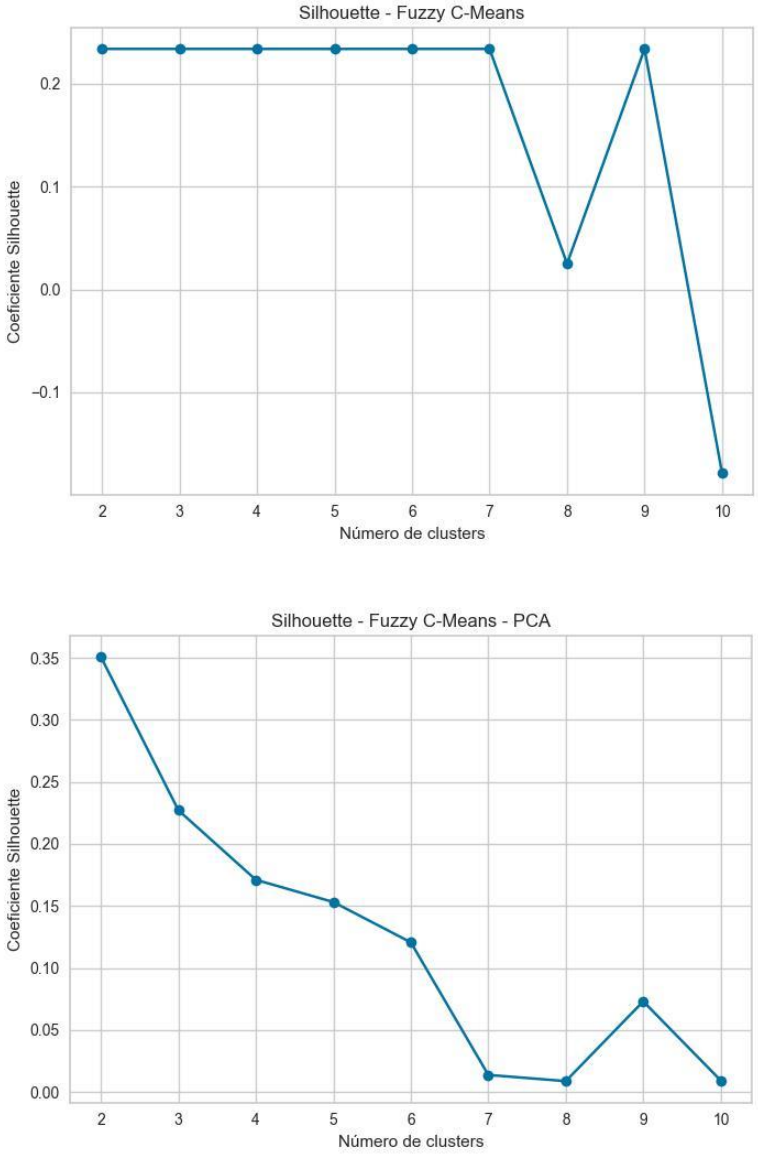


Figura 30: Coeficiente *silhouette* método *Fuzzy C-Means* con y sin aplicación de la técnica PCA.
Fuente: Elaboración propia.

Tabla 9: Número de instancias asignadas en cada *cluster* con el método *Fuzzy C-Means*

Número de *cluster*	***Fuzzy C-Means***	***Fuzzy C-Means - PCA***
0	2	195
1	-	509
2	1048	392
3	-	403
4	555	106

Fuente: Elaboración propia.

En la tabla Tabla 9, se observó que la técnica PCA si presentó mejoría. Para *Fuzzy C-Means*, hay 2 *clusters* vacíos, no se asignaron instancias, y a su vez, el 65 de instancias se asignó a un solo *cluster*. Al contrario, en *Fuzzy C-Means* con PCA, asignó instancias a todos los *clusters* y si bien con son homogéneos, si presentan resultados aceptables en comparación de los métodos anteriores.

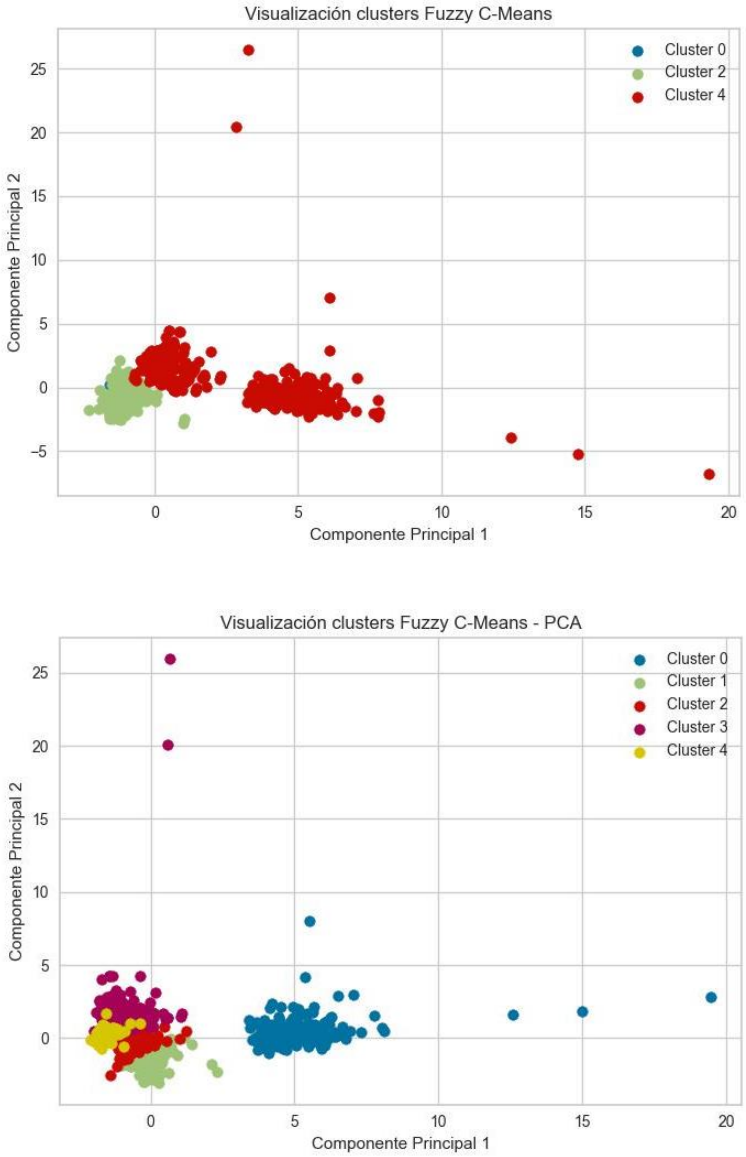


Figura 31: Visualización de *clusters* con método *Fuzzy C-Means* con y sin aplicar técnica PCA. Fuente: Elaboración propia.

En cuanto al coeficiente *silhouette*, se obtuvo lo siguiente:

- Fuzzy C-Means: 0.0744
- Fuzzy C-Means – PCA: 0.2317

Sin embargo, el coeficiente *silhouette* refuta los resultados aceptables al asignar las instancias. De forma gráfica (Figura 31), los *clusters* se observan poco compactos y no definidos.

Comparación y Conclusiones – Clustering

En la Tabla 10 se compararon los métodos de *clustering* utilizados entre sí. Todos tienen un bajo coeficiente *Silhouette*, no superan 0.28. Esto quiere decir que las instancias no fueron clasificadas correctamente y que los grupos no son compactos, ya que las instancias están lejanas del centroide y cerca de los bordes, y debido a esto, es que se consideró que todos los métodos no entregan calidad en el resultado, sugiriendo que la estructura subyacente de los datos puede no ser claramente definida por ninguno de los métodos, o que la naturaleza de los datos no es completamente clusterizable.

Tabla 10: Tabla resumen de los métodos de *clustering* con y sin aplicar la técnica PCA

Método	**Número de Cluster**					**Coeficiente Silhouette**
	0	**1**	**2**	**3**	**4**	
K-Means	980	194	426	3	2	0,2788
K-Means - PCA	974	434	194	2	1	0,2895
Gaussian Mixture	264	180	347	780	54	0,1777
Gaussian Mixture - PCA	240	184	779	368	34	0,1133
Fuzzy C-Means	2	-	1048	-	555	0,0744
Fuzzy C-Means - PCA	195	509	392	403	106	0,2317

Fuente: Elaboración propia.

Además, al comparar la distribución de las instancias, se observa que K-Means generó clusters con tamaños más desiguales en comparación con Gaussian Mixtures y Fuzzy C-Means.

Debido a los antecedentes presentados, se descarta la técnica de *clustering* para este set de datos y no se continuará con la técnica de clasificación propuesta en los objetivos como segunda fase de la técnica de *clustering*.

5.2.2. Métodos de Clasificación para el Crecimiento de Didymo en Sistemas Hídricos de Chile mediante Random Forest y Redes Neuronales

Con el objetivo de desarrollar un modelo de predicción para el crecimiento de Didymo en los sistemas hídricos de Chile, se implementó el método Random Forest para predecir el crecimiento del Didymo en los sistemas hídricos de Chile. El conjunto de datos utilizado fue proporcionado por el Instituto de Fomento Pesquero de Puerto Montt, y se centra en las siguientes clases:

- Crec_alga_Ausent: representa la ausencia de crecimiento del Didymo en el sistema hídrico.
- Crec_alga_Inicial: indica un nivel inicial de crecimiento del Didymo.
- Crec_alga_Mediana: representa un nivel intermedio de crecimiento del Didymo.
- Crec_alga_Alta: indica un nivel alto de crecimiento del Didymo.
- Crec_alga_Muy_Alta: representa el nivel más alto de crecimiento del Didymo.

***Random Forest*:**

***Classification Report*:**

El *Classification Report* (Tabla 11) proporciona métricas detalladas sobre el rendimiento del modelo para cada clase. Sin embargo, los resultados muestran un bajo rendimiento global con precisiones, *recalls* y *f1-scores* muy bajos para todas las clases. La *micro avg* y *weighted avg* confirman que el modelo tiene dificultades para realizar predicciones precisas en todas las clases. La *Accuracy* general es de tan solo 13.08%, lo que indica una baja capacidad del modelo para predecir correctamente las clases.

Tabla 11: Tabla resumen de los resultados de *Classification Report*

	Precision	Recall	f1-score	Support
Crec_alga_Ausent	1.00	0.02	0.04	48
Crec_alga_Inicial	0.48	0.26	0.34	124
Crec_alga_Mediana	0.25	0.03	0.05	78
Crec_alga_Alta	0.00	0.00	0.00	34
Crec_alga_Muy_Alta	1.00	0.04	0.07	28
micro avg	0.47	0.12	0.19	312
macro avg	0.55	0.07	0.10	312
weighted avg	0.50	0.12	0.16	312
samples avg	0.11	0.11	0.11	312

Accuracy:	0.1308411214953271

Fuente: Elaboración propia

Matriz de confusión:

La matriz de confusión (Figura 32) revela que el modelo está teniendo dificultades en la correcta clasificación de las clases. La mayoría de las predicciones se concentran en unas pocas clases, y la confusión entre clases es evidente, como se observa en los valores no nulos fuera de la diagonal principal.

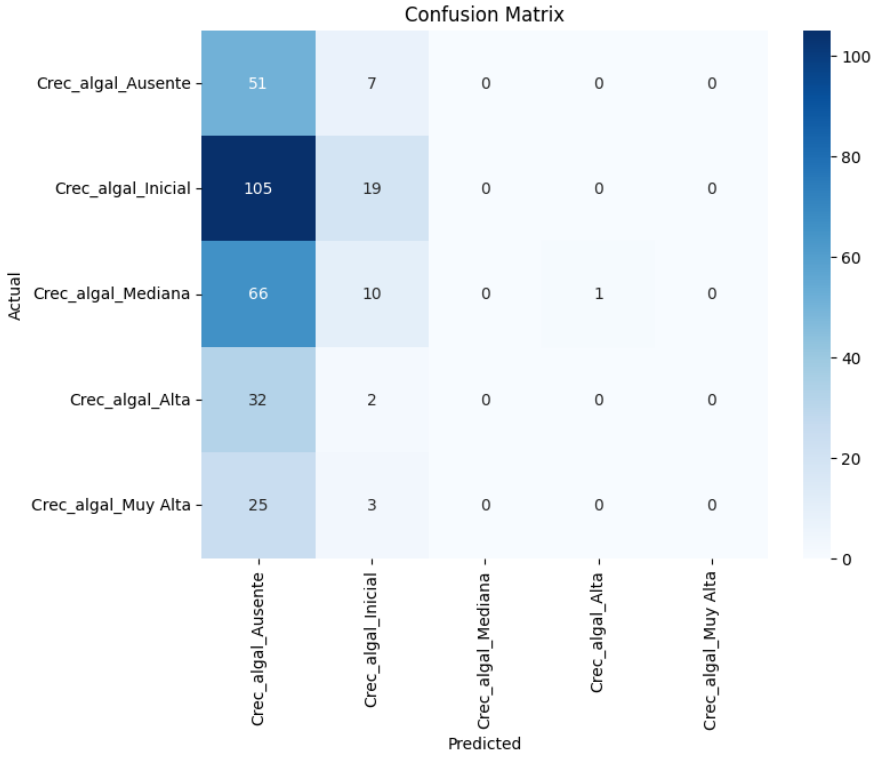


Figura 32: Matriz de Confusión resultante del método *Random Forest*. Fuente: Elaboración propia.

ROC Curve:

El AUC proporciona una métrica adicional de evaluación del modelo. Aunque algunos valores de AUC son superiores a 0.5, lo que sugiere cierta capacidad predictiva, estos valores aún no son lo suficientemente altos como para considerar el modelo como efectivo. Se observa variabilidad en la capacidad predictiva entre las diferentes clases, indicando desafíos específicos en la clasificación de cada categoría (Figura *33*).

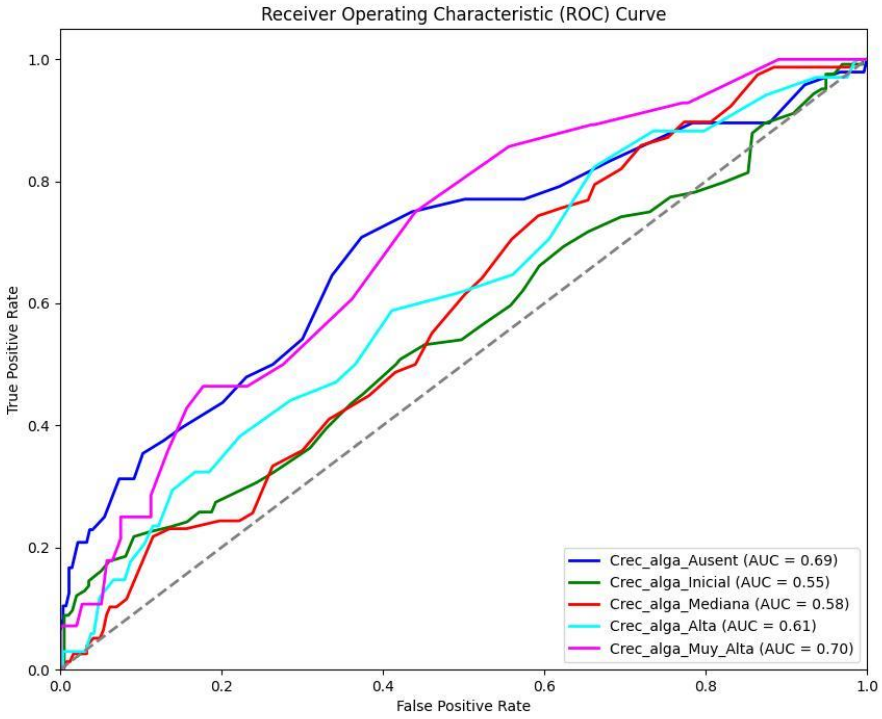


Figura 33: Curva ROC y el valor AUC de cada clase obtenidas del método Random Forest. Fuente: Elaboración propia.

Los resultados del análisis predictivo utilizando *Random Forest* para el crecimiento del Didymo en los sistemas hídricos de Chile indican un rendimiento insatisfactorio del modelo. La baja precisión, *recall* y *f1-score*, junto con una matriz de confusión que muestra una clasificación deficiente.

Redes neuronales

Se realizó un modelo de redes neuronales, el cual obtuvo un 38.63% de exactitud, sin embargo, la curva “loss” terminó su entrenamiento en 1.4294. Este valor indica inestabilidad en el modelo, y si se visualiza el gráfico de curvas “los” y “exactitud”, se observa estabilidad, pero no mejora (Figura 34).

La matriz de confusión no entrega buenos resultado, y solo clasifica la categoría 2 correctamente (Figura 35).

Debido a los antecedentes presentados, se descarta la técnica de *Random Forest* y redes neuronales para este set de datos.

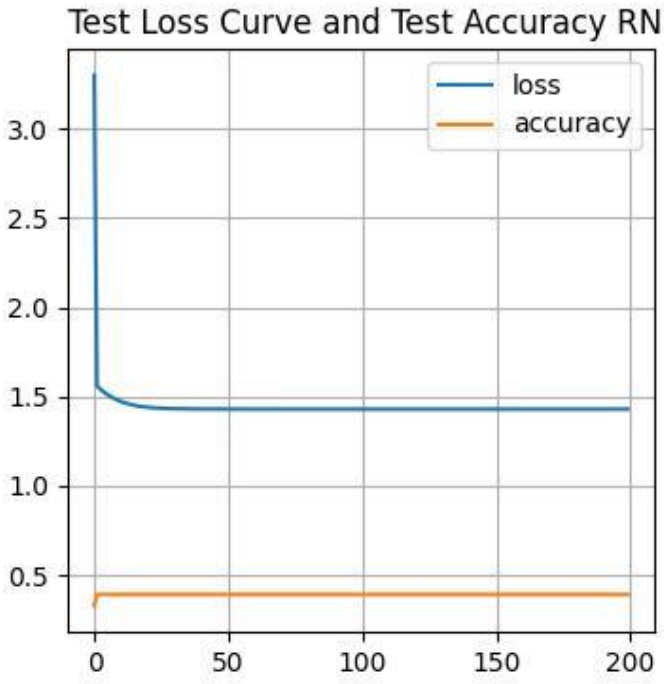


Figura 34: Curvas "loss" y exactitud para modelo de redes neuronales. Fuente: Elaboración propia.

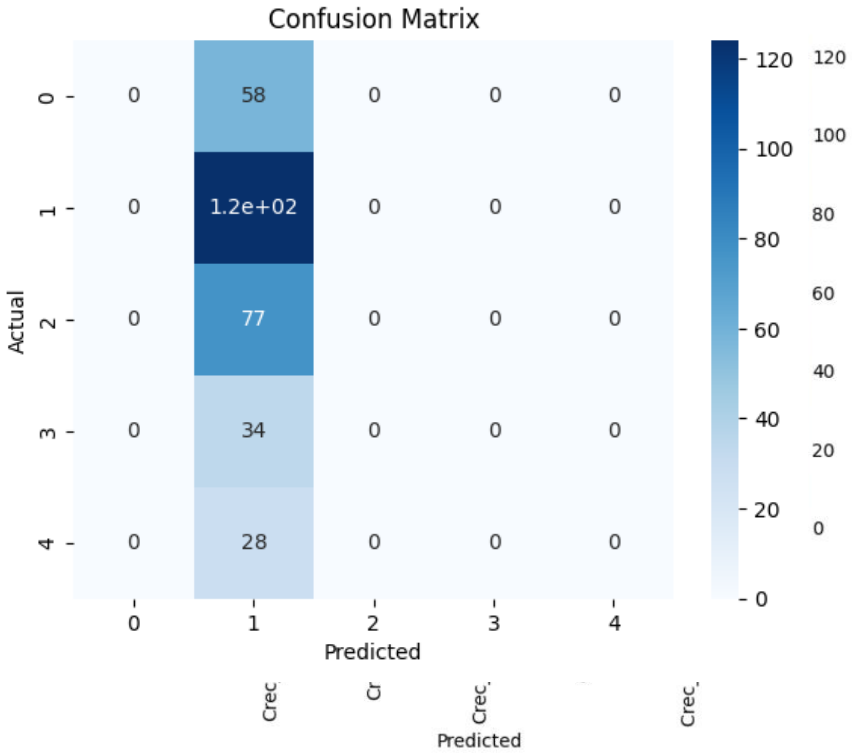


Figura 35: Matriz de confusión para modelo de red neuronal. Fuente: Elaboración propia.

5.2.3. Métodos de Clasificación para el Crecimiento de Didymo en Sistemas Hídricos de Chile mediante Redes Neuronales para Grafos

Los modelos de *clustering* y *ensembled learning* no dieron resultados positivos con el set de datos trabajado, lo que hizo cuestionar la calidad de los datos, por lo que se estudió la correlación entre variables. En la *Figura 36*, se observó que la correlación entre variables es muy baja, predominando valores entre 0 y 0.1.

Estos resultados llevaron a cuestionar los modelos basados en la distancia Euclidiana, en donde las instancias se aglomeran alrededor de un centroide, por lo que se decidió buscar alternativas.

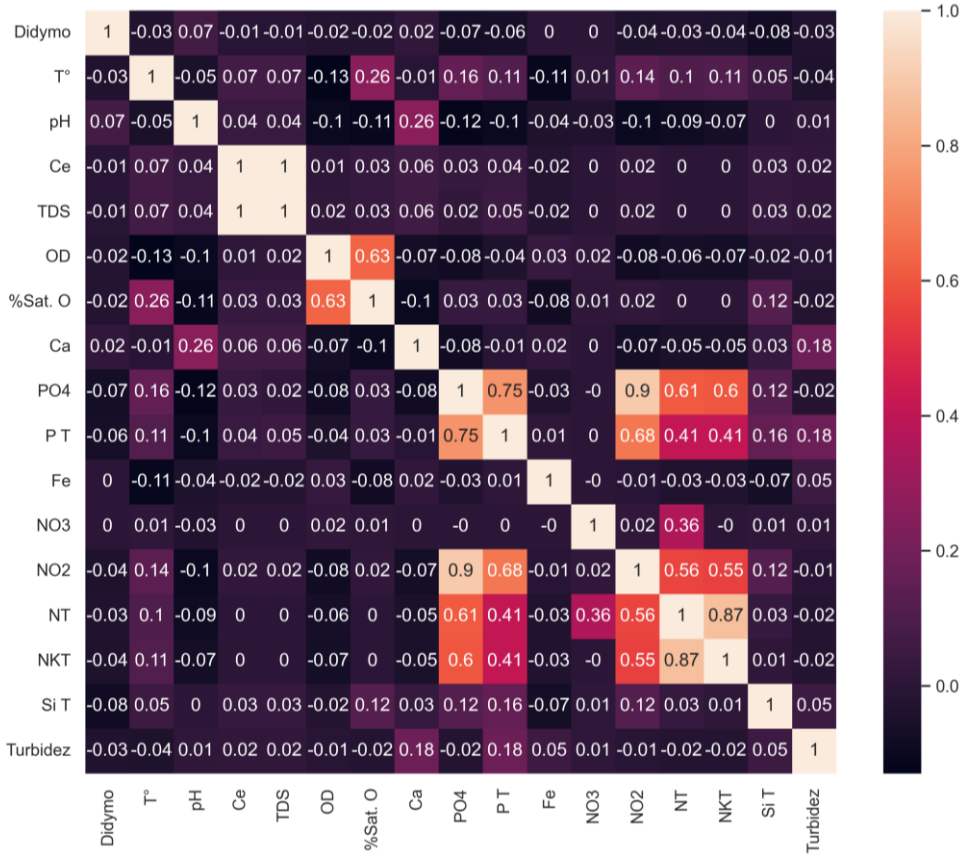


Figura 36: Correlación entre variables del set de datos. Fuente: Elaboración propia.

Se realizaron dos modelos de redes para grafos: redes perceptrón y redes convolucionales. Para ambos métodos se utilizó la misma matriz adyacente de atributos y etiquetas, pero las

relaciones entre nodos fueron diferentes. Para el grafo 1, todos los nodos se relacionaron entre sí, y para grafo 2, se crearon relaciones de la siguiente forma (Tabla 12):

Tabla 12: Relaciones para el grafo

Relaciones	
Variable 1	Variable 2
Crec_algal_Ausente'	%Cob_algal_ausente'
Crec_algal_Inicial'	%Cob_algal_Pequenas colonias'
Crec_algal_Mediana'	%Cob_algal_Mediana'
Crec_algal_Alta'	%Cob_algal_Alta'
Crec_algal_Muy Alta'	%Cob_algal_Muy Alta'
Crec_algal_Ausente'	Didymo = 0
Fosforo Total > 0.112254	

Fuente: Elaboración propia.

Estas relaciones se realizaron para disminuir el número de conexiones y minimizar el gasto computacional, además, de obtener resultados más cercanos a la realidad.
Según la bibliografía, el fósforo es un determinante del crecimiento del didymo (Bravo et al., 2019), por lo que se eligió variable para crear conexiones. El valor Fosforo Total > 0.112254 se calculó mediante la media y la desviación estándar del set de datos, los valores superiores a ese límite corresponden a instancias libres de didymo (en teoría). Además, se relacionó el crecimiento algal con la cobertura algal.

Cada nodo corresponde a un muestreo y los atributos de estos, son los parámetros medidos: 'T°', 'pH', 'Ce', 'TDS', 'OD', '%Sat. O', 'Ca', 'PO4', 'P T', 'Fe', 'NO3', 'NO2', 'NT', 'NKT', 'Si T', 'Turbidez', 'Didymo'.

Las categorías se definieron de la siguiente forma: 'Crec_algal_Ausente', 'Crec_algal_Inicial', 'Crec_algal_Mediana', 'Crec_algal_Alta', 'Crec_algal_Muy Alta'.

Tabla 13: Características de los grafos 1 y 2

	Grafo 1	**Grafo 2**
Número de nodos	1.605	1.605
Número de relaciones	6.436.050	202.046
¿Nodos aislados?	No	Si
¿Self-loops?	No	No
¿Es indirecto?	No	No

Fuente: Elaboración propia.

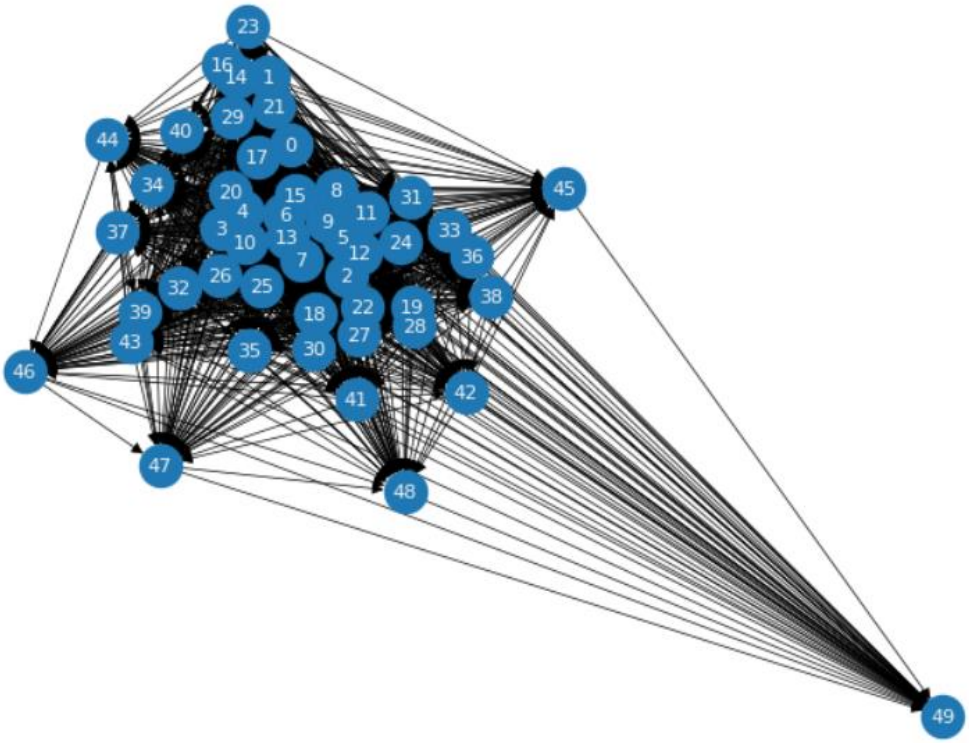


Figura 37: Grafo 1 (se consideraron solo 50 nodos al graficar). Todos los nodos se conectan entre sí. Fuente: Elaboración propia.

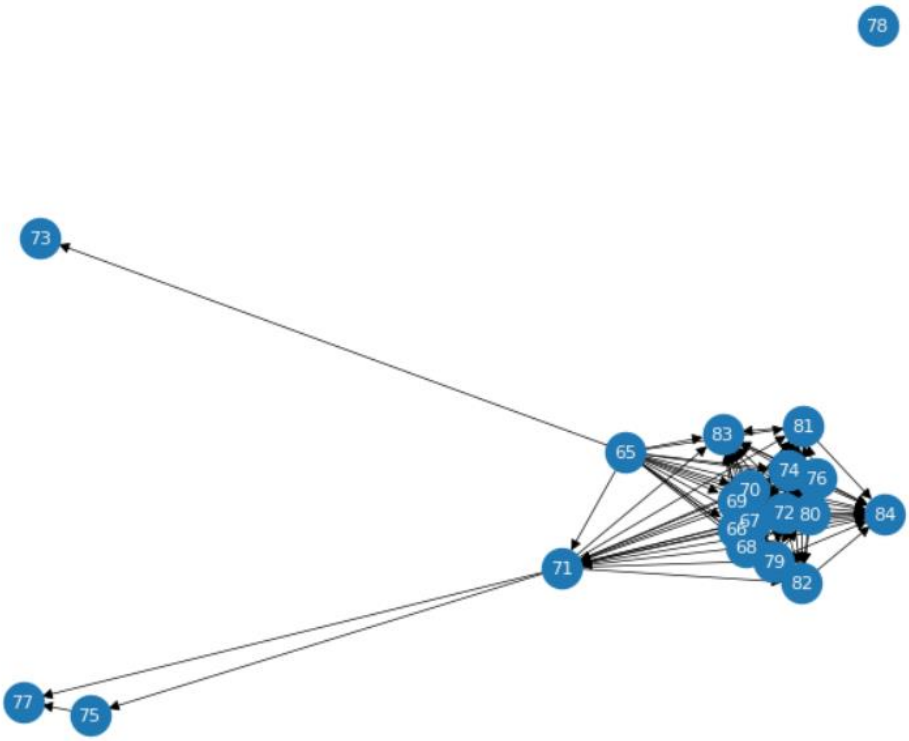


Figura 38: Grafo 2 (se consideraron solo 20 nodos al graficar). Se observan nodos aislados. Fuente: Elaboración propia.

De acuerdo con la Tabla 13, el grafo 1 (Figura 37) cuenta con 1605 nodos y 6.436.050 relaciones, estos debido a que todos los nodos se relacionan entre sí. No hay nodos aislados, *self-loops* y es directo.

El grafo 2 (Figura *38*) cuenta con 1605 nodos, 202.046 relaciones, pero tiene nodos aislados. Esto debido a las relaciones realizadas manualmente. No tiene *self-loops* y también es directo. *Self-loop* indica si el nodo se relaciona consigo mismo, fenómeno que no ocurre en este caso.

Se definieron 6 redes neuronales, 1 red perceptrón, 1 red convolucional y 1 red convolucional normalizada para cada grafo. Las características de estas redes, como número interno de capas, número de neuronas, clases para la clasificación, función de activación y épocas, se detallan en la Tabla 14 y Tabla 15:

Tabla 14: Características de las redes neuronales para el grafo 1

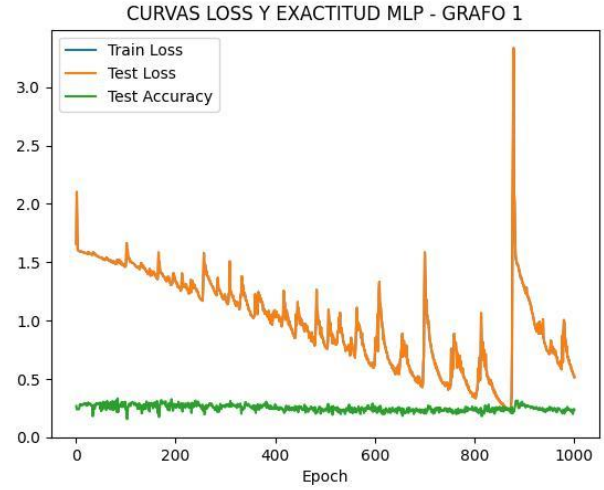
	Grafo 1		
	Red perceptrón (MLP)	**Red Convolucional (GCN)**	**Red Convolucional Normalizada (GCNNorm)**
Número de capas internas	5	3	3
Número de neuronas por capa	96	8	8
Número de clases	5	5	5
Función de activación	relu	relu	relu
Capa de normalización	No	No	Si
Épocas	1000	2000	2000

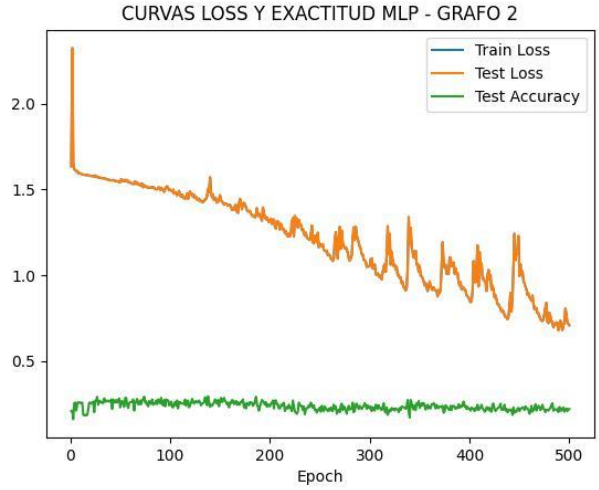
Fuente: Elaboración propia.

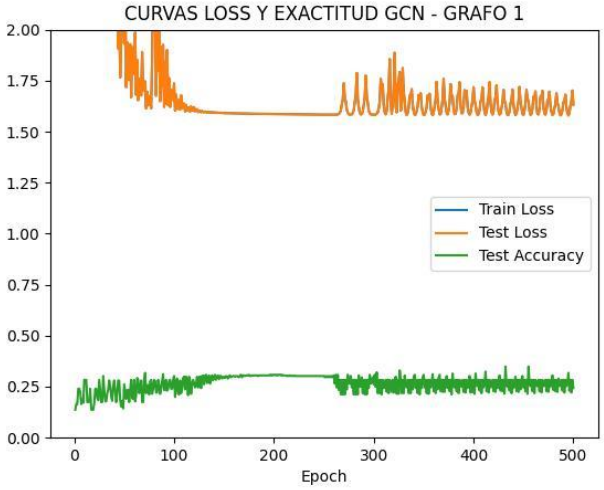
Tabla 15: Características de las redes neuronales para el grafo 2

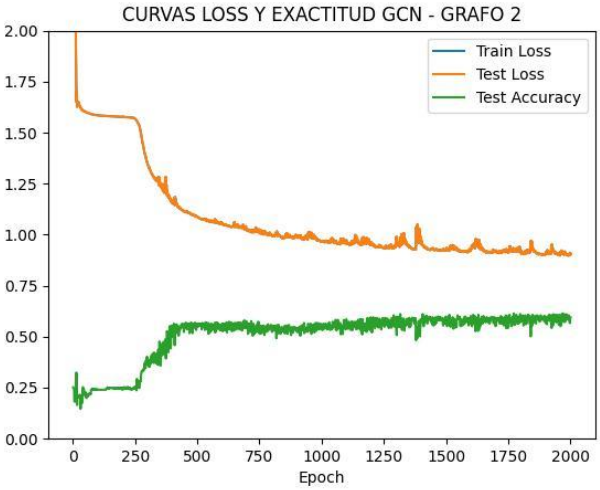
	Grafo 2		
	Red perceptrón (MLP)	**Red Convolucional (GCN)**	**Red Convolucional Normalizada (GCNNorm)**
Número de capas internas	5	3	3
Número de neuronas por capa	96	8	8
Número de clases	5	5	5
Función de activación	relu	relu	relu
Capa de normalización	No	No	Si
Épocas	500	2000	2000

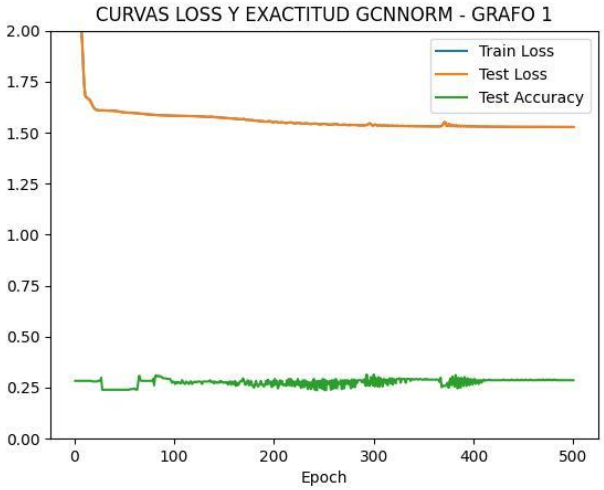
Fuente: Elaboración propia.











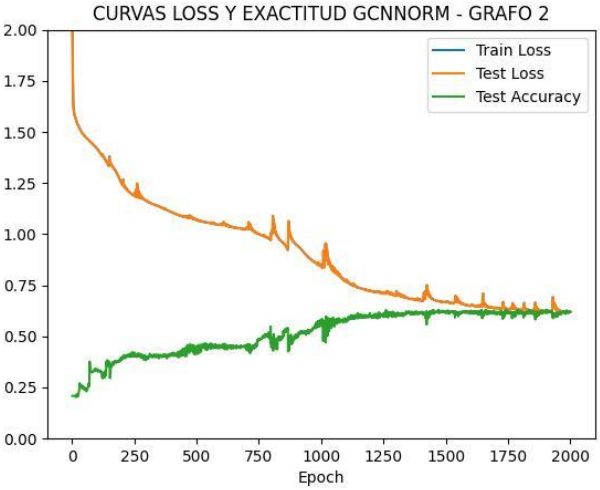


Figura 39: Curvas "loss" y "exactitud" para cada modelo (MLP, GCN y GCNNorm) del grafo 1. Fuente: Elaboración propia.

Figura 40: Curvas "loss" y "exactitud" para cada modelo (MLP, GCN y GCNNorm) del grafo 2. Fuente: Elaboración propia.

Se estudiaron los gráficos con curvas “loss” y “exactitud” para cada grafo. Para el grafo 1, se observa en la Figura 39 que para el modelo MLP la función “loss” es muy inestable pero la función exactitud se muestra estable a lo largo de las 2000 épocas, finalizando con un 23,68%. Para el modelo GCN ocurre lo contrario, ambas funciones son estables, por lo que se podría reducir el número de épocas, sin embargo, la exactitud es del 24.30%, lo cual continúa siendo bajo y la función “los” no decrece de 1.5, lo que indica inestabilidad. En cuanto al modelo GCNNorm, la exactitud es de 28.66%. Las funciones “loss” y exactitud se mantienen constantes, sin embargo “loss” se mantiene sobre 1.5.

Para el modelo MLP del grafo 2, se utilizaron 500 épocas. Se observa que la curva “loss” es muy inestable pero la exactitud se muestra estable y con un leve crecimiento, sin embargo, esta esta es de 21.81%.

Para el modelo GCN del grafo 2, se utilizaron 2000 épocas, la función “loss” disminuye, pero se comporta de forma inestable en ocasiones, para la exactitud es lo contrario, esta aumenta. La exactitud al finalizar las 2000 épocas es de 59.50%.

Para el modelo GCNNorm, se utilizaron 2000 épocas, la función “los” y exactitud se mantienen estables a lo largo de las épocas, lo cual es un indicador para confiar en el modelo. La exactitud es del 61.99%, un resultado mejor que para GCN y solo utilizando 7 segundos extra para el entrenamiento. Estos resultados indicarían que es el mejor modelo y que es necesaria una capa de normalización para la matriz de atributos.

Tabla 16: Resumen de resultados obtenidos de las redes neuronales del grafo 1

	Grafo 1		
	MLP	**GCN**	**GCNNorm**
Épocas	1000	500	500
Exactitud (%)	23.68%	24.30%	28.66%
Tiempo de entrenamiento	13.0 s	18 m, 53.3 s	23 m, 33.9 s

Fuente: Elaboración propia.

Tabla 17: Resumen de resultados obtenidos de las redes neuronales del grafo 2

	Grafo 2		
	MLP	**GCN**	**GCNNorm**
Épocas	1000	2000	2000
Exactitud (%)	21.81%	59.50%	61.99%
Tiempo de entrenamiento	6.1 s	3 min, 30.8 s	3 min, 32.9 s

Fuente: Elaboración propia.

En las Tabla 16 y Tabla 17 se observó la comparación entre los métodos de redes neuronales. El método con la mejor exactitud fue la red convolucional normalizada del grafo 2, con un 64,80%. Esto se debe a menor cantidad de relaciones ingresadas al momento de entrenar el modelo y la normalización de la matriz de atributos.

5.2.4. Resultados para modelos de clasificación

Los resultados obtenidos en los modelos de clasificación realizados se resumen en la Tabla 18:

Tabla 18: Resumen de resultados obtenidos en modelos de clasificación para el set de datos de Didymo

Modelo	Exactitud
Random Forest	13.08%
Red Neuronal	38.63%
MLP – Grafo 1	23.68%
GCN – Grafo 1	24.30%
GCN normalizado – Grafo 1	28.66%
MLP – Grafo 2	21.81%
GCN – Grafo 2	59.50%
GCN normalizado – Grafo 2	61.99%

Fuente: Elaboración propia

Como se observa en la Tabla 18, el modelo de red neuronal convolucional normalizada para el grafo 2 obtuvo el porcentaje de exactitud más alto, con un 61.99%, por lo que este modelo se elige el mejor modelo de clasificación para este set de datos.

El grafo 2 contiene 16 parámetros químicos y la cantidad de células de Didymo presentes en el sitio de toma de muestra como atributos, las relaciones se realizaron según la presencia de Didymo y la cobertura algal presente en el sitio de muestreo, además de la concentración de fosforo, ya que este es el principal parámetro involucrado en el crecimiento del Didymo.

En un estudio realizado por Hix & Murdock (Hix & Murdock, 2019), se realizó un modelo de clasificación Random Forest para predecir la ausencia de Didymo o la presencia de células o manto, para lo cual utilizaron la concentración de células detectadas por cm^2 y el

porcentaje de cobertura (expresado en %), sin embargo, no consideraron parámetros químicos. Con un set de datos de 70 instancias, obtuvieron una exactitud del 81.80%

En el modelo que se propuso, se consideraron parámetros químicos, se obtuvo un porcentaje mejor de exactitud, sin embargo, la complejidad de los datos y la cantidad de atributos considerados, abren la posibilidad de extender este modelo de predicción de ausencia o presencia de Didymo a cualquier época del año (invierno o verano), sin poner en riesgo a la persona que toma la muestra, al tener que ingresar a ríos con caudales altos o con condiciones meteorológicas extremas, ya que solo es necesaria una muestra de agua.

6. Conclusiones y trabajo futuro

A continuación, se presentarán las conclusiones y trabajo futuro para el presente trabajo de fin de máster, el cual involucró un arduo trabajo investigación y de aplicación de técnicas de análisis de datos e inteligencia artificial.

6.1. Conclusiones

La propuesta del trabajo de fin de máster consistía en realizar una nueva similitud de los datos a través de *clustering* y la posterior clasificación de estos para conocer la ausencia o presencia y en que magnitud se encuentra el Dydimo a través del muestreo de parámetros químicos *in situ* en ríos del sur de Chile, sin embargo, no fue posible realizarlo, pero se propuso un modelo de grafos.

El origen de los datos trabajados es muy complicado, no existe relación entre las mediciones ya que se trata de datos vivos, es decir, datos de la naturaleza, volátiles e impredecibles. Es por ello, que técnicas como *clustering*, no son posibles de aplicar, ya que, al basarse en modelos Euclidianos, la convierte en una herramienta rígida, lo mismo ocurre con *Random Forest*.

Para resolver este problema se propuso un modelo de grafos, en donde cada nodo contenía las mediciones de parámetros químicos y las conexiones correspondían al estado del Dydimo. Este modelo entrego resultados aceptables y mejorables (en comparación de *clustering* y *Random Forest*) y dieron lugar al cumplimiento de los objetivos específicos propuestos:

- **Objetivo específico 1** (Limpiar, analizar, interpretar y normalizar a través de un análisis exploratorio de datos (EDA) los datos registrados): Se eliminaron caracteres especiales y se completaron los datos faltantes del set de datos teniendo en cuenta los sitios de muestreos específicos (ríos, cuencas o subcuencas), para variables numéricas se utilizó la media y para variables binarias se utilizó la moda. Se hizo un estudio de correlación entre variables para la toma de decisiones de los modelos a utilizar.
- **Objetivo específico 2** (Clasificar datos registrados): Esta clasificación inicialmente se realizaría a través de la técnica de *clustering*, sin embargo, cuando se percató la

inviabilidad de la herramienta en este tipo de datos, se realizó a través de un modelo de clasificación de grafos.

- **Objetivo específico 3** (Aplicar y comparar modelos de aprendizaje supervisado como *Ensembled Learning* y Redes Neuronales): Se realizaron estos modelos, sin embargo, al no entregar buenos resultados, se decidió no continuar y utilizar otras técnicas, como redes neuronales no convolucionales para grafos.
- **Objetivo específico 4** (Entrenar y evaluar los modelos implementados para comparar métricas de rendimiento): El entrenamiento y evaluación se realizó en un modelo de clasificación de grafos con redes neuronales convolucionales para lo cual, en primer lugar, se diseñó un modelo de grafo de acuerdo con los parámetros químicos determinantes para el crecimiento de Didymo, como el fósforo. Se obtuvieron resultados numéricos aceptables, para ello, se analizaron gráficas de funciones que entregan información sobre los modelos y valores numéricos.
- **Objetivo específico 5** (Generar un informe con los hallazgos conseguidos): Se registraron los modelos exitosos y no exitosos y se compararon entre sí, evaluando métricas y mostrando resultados numéricos y visuales.

Por lo tanto, el objetivo general se cumplió, no a través de las técnicas propuestas inicialmente, pero si con un modelo de clasificación de red convolucional para grafos.

6.2. Líneas de trabajo futuro

El resultado obtenido en el presente trabajo de fin de máster da cabida a la utilización de este tipo de técnicas en áreas de investigación lejanas al análisis de datos o inteligencia artificial, como lo es el monitoreo ambiental. Actualmente los datos recopilados por centros de investigación, empresas u otro tipo de organización, no están siendo trabajados correctamente, ya que, si bien se están trabajando en bases de datos, no se está obteniendo conocimiento. Dicho esto, se propone como trabajo futuro lo siguiente:

- Recopilación de un mayor volumen de datos, de otros centros de investigación o de otros países para entrenar el modelo de clasificación y mejorar la exactitud del modelo para la clasificación
- Ver la posibilidad de agregar otro tipo de parámetros químicos o parámetros físicos al modelo

- Para el grafo, estudiar relaciones entre nodos, como la relación del fósforo, y acotarla de forma más precisa y también, seleccionar otros parámetros y crear más relaciones
- Seguir profundizando con la red neuronal no convolucional, estudiando otro tipo de capas (como la de normalización)
- Extender este modelo a otras especies de algas o microorganismos, para prevenir desastres naturales en ríos, lagos o mares de Chile o de otros países

7. Bibliografía

AMAKAIK Consultoría Ambiental. (2014). *Evaluación de Didymosphenia geminata (Didymo) en cuerpos de agua de la zona centro-sur*.

Añón, D., & Albariño, R. (2020). Efecto del establecimiento del alga invasora Didymosphenia geminata sobre la abundancia de macrocrustáceos en el Río Limay superior, Patagonia, Argentina. *Biología Acuática*, *34*, 006. https://doi.org/10.24215/16684869e006

Asif, N. A., Sarker, Y., Chakrabortty, R. K., Ryan, M. J., Ahamed, M. H., Saha, D. K., Badal, F. R., Das, S. K., Ali, M. F., Moyeen, S. I., Islam, M. R., & Tasneem, Z. (2021). Graph Neural Network: A Comprehensive Review on Non-Euclidean Space. *IEEE Access*, *9*, 60588–60606. https://doi.org/10.1109/ACCESS.2021.3071274

Beeby, J. (2012). *Water quality and survivability of Didymosphenia Geminata* [Tesis de maestría, Colorado State University]. http://hdl.handle.net/10217/73549

Betancurt, R. F., Baffico Guadalupe Beamud, G. S., Baffico, G. D., Biología, en, Guadalupe Beamud, S., en Biología, D., & Biológicas, I. (2016). Alga Didymo: una pequeña gran invasora. *Desde La Patagonia Difundiendo Saberes*, *14*(23), 28–34.

Blanco, S., & Ector, L. (2009). Distribution, ecology and nuisance effects of the freshwater invasive diatom Didymosphenia geminata (lyngbye) M. Schmidt: A literature review. In *Nova Hedwigia* (Vol. 88, Issues 3–4, pp. 347–422). https://doi.org/10.1127/0029-5035/2009/0088-0347

Bothwell, M. L., Lynch, D. R., Wright, H., & Deniseger, J. (2009). On the Boots of Fishermen: The History of Didymo Blooms on Vancouver Island, British Columbia. *Fisheries*, *34*(8), 382–388. https://doi.org/10.1577/1548-8446-34.8.382

Bothwell, M. L., Taylor, B. W., & Kilroy, C. (2014). The Didymo story: The role of low dissolved phosphorus in the formation of Didymosphenia geminata blooms. *Diatom Research*, *29*(3), 229–236. https://doi.org/10.1080/0269249X.2014.889041

Bravo, S., Whelan, K., Sambra, K., Silva, M. T., Ponce, N., & Campos, P. (2019). Comparative analysis of two rivers infected with Didymosphenia geminata in southern Chile. *Latin American Journal of Aquatic Research*, *47*(4), 665–676. https://doi.org/10.3856/vol47-issue4-fulltext-8

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2016). *Geometric deep learning: going beyond Euclidean data*. https://doi.org/10.1109/MSP.2017.2693418

Burkholder, J. M. (2009). Harmful Algal Blooms. In *Encyclopedia of Inland Waters* (pp. 264–285). http://www.keweenawalgae.mtu.edu

Cameron, N. G. (2013). Diatoms. In *Encyclopedia of Quaternary Science: Second Edition* (pp. 522–525). Elsevier Inc. https://doi.org/10.1016/B978-0-444-53643-3.00359-9

Capdevila-Argüelles, L., Zilletti, B., & Suárez, V. (2011). *Manual de las especies exóticas invasoras de los ríos y riberas de la cuenca hidrográfica del Duero*. http://opencage.info:

Chollet, F. (2018). *Deep Learning with Python* (1st ed.). Manning Publications Co.

Díaz, C., Molina, X., & Montecino, V. (2011). Manual para el Monitoreo e Identificación de la Microalga Bentónica Didymosphenia geminata. In *Subsecretaría de Pesca Gobierno de Chile*.

Díaz, C., Salcedo, F., Muñoz, S., Bozo, D., Santibáñez, C., Fernández, C., Aguilar, F., Labra, F., Moreno, R., Cayupe, B., & Ehrenfeld, N. (2017). *PROYECTO FIPA N°2015-04: Monitoreo de la especie plaga Didymosphenia geminata en cuerpos de agua de la zona centro sur austral*.

Diaz, C., Salcedo, F., Olivares, M., & Maidana, N. (2016). *Manual para el monitoreo e identificación de la microalga bentónica Didymosphenia geminata*.

Esse, C., Fustos, I., González, K., Aguayo, C., Encina-Montoya, F., Figueroa, D., Lara, G., & Navarro, C. (2018). Spectral characterization of didymosphenia geminata under laboratory conditions: Bases for a monitoring and early warning system in river systems of south Central Chile. *Management of Biological Invasions*, *9*(2), 85–90. https://doi.org/10.3391/mbi.2018.9.2.02

Figueroa, F., Pedreros, P., Bravo, G., & Urrutia, R. (2021). *Uso de especies invasoras de agua dulce: Una potencial estrategia de economía circular*. https://www.researchgate.net/publication/351244535

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (N. Tache, Ed.; 2nd ed.). O'Reilly Media, Inc.

Gouvernement du Québec. (2008). *What is Didymo and how can we prevent it from spreading in our rivers?* Ministère du développement durable, de l'environnement et des parcs Québec.

Hanna Instruments. (2023). *General Catalog Volumen 37*. www.hannainst.com

Hix, L. A., & Murdock, J. N. (2019). Didymosphenia geminata habitat requirements are unique and variable for cell establishment and mat accumulation. *Hydrobiologia*, *828*(1), 147–164. https://doi.org/10.1007/s10750-018-3809-3

Iturrieta, M. (2016). *Análisis exploratorio de la relación entre el régimen de caudales y la abundancia de Didymosphenia geminata en los ríos de la zona centro sur de chile*.

Jellyman, P. G., Clearwater, S. J., Biggs, B. J. F., Blair, N., Bremner, D. C., Clayton, J. S., Davey, A., Gretz, M. R., Hickey, C., & Kilroy, C. (2006). *Didymosphenia geminata experimental control trials: Stage One (screening of biocides and stalk disruption agents) and Stage Two Phase One (biocide testing)*. www.niwa.co.nz

Kawecka, B., & Sanecki, J. (2003). Didymosphenia geminata in running waters of southern Poland-symptoms of change in water quality? In *Hydrobiologia* (Vol. 495).

Kilroy, C. (2004). *A new alien diatom, Didymosphenia geminata (Lyngbye) Schmidt: its biology, distribution, effects and potential risks for New Zealand fresh waters*. www.niwa.co.nz

Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *Conference Paper at ICLR 2017*. http://arxiv.org/abs/1609.02907

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. In *Secondary Analysis of Electronic Health Records* (pp. 185–203). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_15

Labib, W., El-Dahhar, A. A., Shahin, S. A., Ismail, M. M., Hosny, S., & Diab, M. H. (2023). Water quality indices as tools for assessment of the Eastern Harbor's water status (Alexandria, Egypt). *SN Applied Sciences*, *5*(3). https://doi.org/10.1007/s42452-023-05304-z

Lamaro Anabel A., Pisonero Juliana, Uyua Noelia, Sastre Viviana, Santinelli Norma, Norma, S., Julieta Muñiz, & Sala, S. E. (2019). *Distribución de la diatomea invasora Didymosphenia geminata (Bacillariophyceae) en cuerpos de agua patagónicos de Argentina.* http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-23722019000200002&lng=es&tlng=es.

Larned, S., Arscott, D., Blair, N., Jarvie, B., Jellyman, D., Lister, K., Schallenberg, M., Sutherland, S., Vopel, K., & Wilcock, B. (2007). *Ecological studies of Didymosphenia geminata in New Zealand, 2006-2007 MAF Biosecurity New Zealand*. www.niwa.co.nz

Manning, C. (2020). *Brief Definitions of Key Terms in AI*. Human-Centered Artificial Intelligence - Stanford University; Stanford University. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Maurya, S. K., Liu, X., & Murata, T. (2023). Feature selection: Key to enhance node classification with graph neural networks. *CAAI Transactions on Intelligence Technology*, *8*(1), 14–28. https://doi.org/10.1049/cit2.12166

Ministerio de Asuntos Económicos y Transformación Digital. (2021). *Guía práctica de introducción al Análisis Exploratorio de Datos*. https://datos.gob.es/sites/default/files/doc/file/analisis_exploratorio_de_datos_2021_v6_0.pdf

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python* (D. Schanafelt, Ed.; 1st ed.). O'Reilly Media, Inc.

Needham, M., & Hodler, A. (2019). *Graph Algorithms. Practical Examples in Apache Spark & Neo4j* (J. Bleiel, Ed.; 1ra ed.). O'Reilly.

Oyanedel, A., Ordóñez, P., & Rojas, R. (2022). *Informe Final: Convenio Desempeño 2021-2022: Monitoreo de la especie plaga Didymosphenia geminata en cuerpos de agua de la zona centro, sur y austral de Chile, Etapa VI, 2021-2022*.

Pajankar, A., & Joshi, A. (2022). Hands-on machine learning with python: Implement neural network solutions with scikit-learn and PyTorch. In *Hands-on Machine Learning with Python: Implement Neural Network Solutions with Scikit-learn and PyTorch*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-7921-2

Raschka, S., & Mirjalili, V. (2019). *Python machine learning : machine learning and deep learning with python, scikit-learn, and tensorflow 2* (J. Malysiak, Ed.; 3rd ed.). Packt Publishing Ltd.

Ren, M., & Mackay, M. (2019). *CSC 411: Introduction to Machine Learning*. University of Toronto. https://www.cs.toronto.edu/~mren/teach/csc411_19s/lec/lec16.pdf

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases* (M. Beaugureau, Ed.; 2nd ed.). O'Reilly.

Sabater, S. (2009). Diatoms. In *Encyclopedia of Inland Waters* (pp. 149–156).

Salas, N., Lengua, R., Becerra, E., Bazán, D., Santome, S., & Córdova, C. (2014). Cuantificación de carbohidratos, polifenoles y sulfatos en extractos de macroalgas promisorias para la acuicultura Quantification of carbohydrates, polyphenols and sulfate on selected macroalgae promising for aquaculture. *Rev. Per. Quím. Ing. Quím*, *17*(2), 35–40.

Salvo, J., & Oyanedel, A. (2019). Community signals of the effect of didymosphenia geminata (Lyngbye) M. Schmidt on benthic diatom communities in Chilean rivers. *Revista Chilena de Historia Natural*, *92*(1), 1–12. https://doi.org/10.1186/s40693-019-0084-2

Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. (2021). A Gentle Introduction to Graph Neural Networks. *Distill*, *6*(8). https://doi.org/10.23915/distill.00033

Segura, P. (2011). A slimy invader blooms in the rivers of Patagonia. In *Science* (Vol. 331, Issue 6013, p. 18). https://doi.org/10.1126/science.331.6013.18

Seltman, H. J. (2018). *Experimental Design and Analysis*. Carnegie Mellon University.

Sheath, R. G., & Wehr, J. D. (2015). Introduction to the Freshwater Algae. In *Freshwater Algae of North America: Ecology and Classification* (pp. 1–11). Elsevier Inc. https://doi.org/10.1016/B978-0-12-385876-4.00001-3

Sterrenburg, F., Gordon, R., Tiffany, M. A., & Nagy, S. S. (2007). Diatoms: Living in a Constructal Environment. In *Algae and Cyanobacteria in Extreme Environments* (pp. 141–172).

SUBPESCA. (2010). *Informe Técnico N°1681/2010. Presencia de Didymosphenia geminata en río Espolón y río Futaleufú, Región de Los Lagos*.

SUBPESCA. (2021). *Resolución Exenta N°719-2021: Declara área de plaga y de riesgo de plaga que indica en cuerpos de agua que señala en materia de acuicultura.*

SUBPESCA. (2022). *Resolución Exenta N°1854-2022: Declara área de plaga y de riesgo de plaga que indica en cuerpos de agua que señala en materia de acuicultura*.

Sundareshwar, P. V., Upadhayay, S., Abessa, M., Honomichl, S., Berdanier, B., Spaulding, S. A., Sandvik, C., & Trennepohl, A. (2011). Didymosphenia geminata : Algal blooms in oligotrophic streams and rivers. *Geophysical Research Letters*, *38*(10), n/a-n/a. https://doi.org/10.1029/2010gl046599

Sutherland, S., Rodway, M., Kilroy, C., Jarvie, B., & Hughes, G. (2007). *The survival of Didymosphenia geminata in three rivers and associated groundwater fed tributaries in the South Island of New Zealand*.

Tapia, N. (2012). *Caracterización física y química de la cuenca hidrográfica del Yelcho y su relación con la presencia de la especie plaga Didymosphenia Geminata*. Universidad de Chile.

Valente De Oliveira, J., & Pedrycz, W. (2007). *Advances in Fuzzy Clustering and its Applications* (J. Valente De Oliveira & W. Pedrycz, Eds.).

Watson, S. B., Whitton, B. A., Higgins, S. N., Paerl, H. W., Brooks, B. W., & Wehr, J. D. (2015). Harmful Algal Blooms. In *Freshwater Algae of North America: Ecology and Classification* (pp. 873–920). Elsevier Inc. https://doi.org/10.1016/B978-0-12-385876-4.00020-7

Whitton, B. A., Ellwood, N. T. W., & Kawecka, B. (2009). Biology of the freshwater diatom Didymosphenia: a review. *Hydrobiologia*, *630*(1), 1–37. https://doi.org/10.1007/s10750-009-9753-5

Wu, Y. (2017). Indicators for Monitoring Aquatic Ecosystem. In *Periphyton* (pp. 71–106). Elsevier. https://doi.org/10.1016/b978-0-12-801077-8.00003-x

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). *A Comprehensive Survey on Graph Neural Networks*. https://doi.org/10.1109/TNNLS.2020.2978386

Zhou, Z. H. (2021). Machine Learning. In *Machine Learning*. Springer Nature. https://doi.org/10.1007/978-981-15-1967-3

Anexos

Anexo I. Código Implementado

Repositorio general código implementado: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/tree/main

Notebook para análisis exploratorio de datos: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/EDA%20TFM.ipynb

Notebook *clustering*:

- *Clustering* para el set de datos: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Clustering-data.ipynb
- *Clustering* para el set de datos con técnica PCA: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Clustering-pca.ipynb

Notebook *Random Forest*: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Random%20Forest%20TFM.ipynb

Notebook Red Neuronal: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Neural%20Network%20TFM.ipynb

Notebook grafo 1: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Grafo%20-1%20TFM.ipynb

Notebook grafo 2: https://github.com/EstefaniaAracena/TFM---Estefania-Aracena-Lenin-Medina/blob/main/Grafo%20-2%20TFM.ipynb

Anexo II. Solicitud de datos

Subsecretaría de Pesca y Acuicultura
Gobierno de Chile

(UPCG) ORD Nº: 00142/2023

ANT.: Consulta Nº AH002T-0005906

MAT.: Derivación de solicitud de información pública que indica N°AH002T-0005906

Valparaíso, 16/10/2023

DE: JAVIER ANDRES RIVERA VERGARA
SUBSECRETARIO (S)
SUBSECRETARIA DE PESCA Y ACUICULTURA

A: DIRECTOR EJECUTIVO INSTITUTO DE FOMENTO PESQUERO
INSTITUTO DE FOMENTO PESQUERO

Por medio de la presente, informamos que hemos recibido la solicitud de información de la ciudadana Estefanía Aracena, quien solicita lo siguiente: "...los datos obtenidos de muestreos químicos, físicos y microbio lógicos (o todos los que tengan), en relación a los proyectos de Didymo realizados por el Instituto de Fomento Pesquero IFOP, desde el año 2010 en adelante (o los años disponibles).

Estos datos serán utilizados en la realización de una tesis de magister: "Este trabajo de fin de máster consiste en un Sistema de alerta ante la presencia de Didymosphenia Geminata en agua dulce con técnicas de inteligencia artificial, el cuál será presentado en el Máster Universitario en Análisis y Visualización de Datos Masivos de la Universidad de La Rioja. Para esto se requirieran los datos que han recolectado desde el comienzo del proyecto y se deben incluir datos de parámetros físico-químicos, nutrientes, fechas y ubicaciones, además de todo dato que sea determinante en el estudio."

En virtud de lo establecido en el artículo 13° de la Ley N°20.285 sobre Acceso a la Información Pública, solicitamos a Ud., pueda hacer entrega de la información requerida. Finalmente, solicitamos hacer llegar la respuesta en forma directa a su correo electrónico e.aracenavallejos@gmail.com

SALUDA ATENTAMENTE A UD.

ESTE DOCUMENTO CUENTA CON FIRMA ELECTRÓNICA AVANZADA
Subsecretaría de Pesca y Acuicultura

JAVIER ANDRES RIVERA VERGARA
SUBSECRETARIO (S)
SUBSECRETARIA DE PESCA Y ACUICULTURA

MGP/MCM